Effective Feature Space Reduction with Imbalanced Data for Semantic Concept Detection

Lin Lin, Guy Ravitz, Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
l.lin2@umiami.edu, {ravitz,shyu}@miami.edu

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

Abstract

Semantic understanding of multimedia content has become a very popular research topic in recent years. Semantic concept detection algorithms face many challenges such as the semantic gap and imbalance data, among others. In this paper, we propose a novel algorithm using multiple correspondence analysis (MCA) to discover the correlation between features and classes to reduce the feature space and to bridge the semantic gap. Moreover, the proposed algorithm is able to explore the correlation between items (i.e., feature-value pairs generated for each of the features) and classes which expands its ability to handle imbalance data sets. To evaluate the proposed algorithm, we compare its performance on semantic concept detection with several existing feature selection methods under various well-known classifiers using some of the concepts and benchmark data available from the TRECVID project. The results demonstrate that our proposed algorithm achieves promising performance, and it performs significantly better than those feature selection methods in the comparison for the imbalanced data sets.

1. Introduction

Multimedia retrieval has a long history and many approaches have been developed to manage and query diverse data types in the computer systems [17]. Semantic understanding of multimedia content is the final frontier in multimedia information retrieval. One of the fundamental challenges in semantic understanding of multimedia content is *semantic concept detection*. The desired concepts to be detected could be the existence of an entity such as faces, trees, etc, or of a more descriptive meaning such as weather, sports, and more. Content-based concept detection applications use low-level features, such as visual features, text-based features, audio features, motion features, and other meta data to determine the semantic meaning from the multimedia data. As previously mentioned, two issues, namely semantic gap and imbalanced data, have been identified as the main obstacles that any system faces when attempting to understand the semantics of multimedia content. Recently many researchers have directed their efforts towards the development of machine learning algorithms that will have the capability to bridge the so-called semantic gap. Schneiderman and Kanade [16] proposed a system for component-based face detection using statistics of parts. A framework for a multimodal video event detection system which combined the analysis of both speech recognition and video annotations was developed [3]. Chen et al. [4] proposed a framework using both multimodal analysis and temporal analysis to offer strong generality and extensibility with the capability of exploring representative event patterns. In [5], the authors proposed a user-centered semantic event retrieval framework which incorporated the Hierarchical Markov Model Mediator mechanism.

In theory, any supervised learning algorithm could be used for semantic concept detection. However, that is under the assumption that the distribution of positive and negative data is balanced in the training data. In fact, this may not always be true in real multimedia databases, which usually only include a small collection of positive instances for some semantic concepts. When the data set is imbalanced, many machine learning algorithms have problems, and the prediction performance can significantly decrease [11]. The two most common sampling schemes which are currently used to adapt machine learning algorithms to imbalanced data sets are called over-sampling and under-sampling. The first adapts the algorithm to the imbalanced data by duplicating the positive data and increasing the frequency of the positive class in the training set; while the second scheme

does so by discarding some negative data and by that it balances the frequency of the positive and negative classes in the training set. Some existing solutions which have been proposed to handle the imbalanced data set problem are the analysis of the relationships between the class distribution of a fixed size training data [18], an approach combining different expressions of the resampling method [8]. Finally, in our previous work [13], we proposed a pre-filtering architecture to prune the negative instances using association rule mining.

Feature selection is one of the most frequently used technique in data pre-processing to remove redundant, irrelevant, and noisy data. By reducing the feature space, the efficiency, accuracy, and comprehensibility of the algorithm could be improved. Ideally, using feature selection would make it possible for the system to choose a feature subset from the original feature set, which best represents the target semantic concepts. The performance of a feature subset is measured by an evaluation criterion which is selected based on the evaluation model that is used. The three main evaluation models that are used for feature selection are the filter model [9][22], the wrapper model [7][12], and the hybrid model [6][21]. The filter model uses the independent evaluation functions, the wrapper model uses the performance of one pre-determined algorithm as the dependent evaluation criterion, and the hybrid model takes advantage of the two models by using different evaluation criteria in different search stages. Both supervised feature selection algorithms (i.e., the feature selection used for classification with labeled data) and the unsupervised feature selection algorithms (i.e., the feature selection used for clustering with unlabeled data) have been developed [14].

In this paper, we propose a novel feature selection algorithm using Multiple Correspondence Analysis (MCA) [15] to evaluate the extracted low-level features. Using the best feature subset captured by MCA, we compare the performance of the semantic concept detection between the proposed framework with the performance of several other existing feature selection algorithms using some of the concepts and benchmark data from TRECVID 2007 [1] under various well-known classifiers, such as the Decision Tree (C4.5), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), and Naive Bayes (Bayes). Overall, the proposed framework performs better than other feature selection methods over all five classifiers, and performs significantly better with imbalanced data sets.

This paper is organized as follows. In Section 2, the different technologies of filter model based feature selection and classification methods are introduced. Section 3 presents the details of the proposed feature selection algorithm. Section 4 discusses our experimental results, and the conclusion is provided in Section 5.

2 Relevant Technologies

In this section, we introduce those algorithms that we used in the performance comparison. There are many feature space reduction algorithms and classification algorithms available in the literature. Here, several most popular algorithms are selected.

2.1 Feature Space Reduction Algorithms

As previously mentioned, there are supervised and unsupervised feature selection methods which are used to reduce the high-dimensional data sets. For the supervised filter-based feature selection algorithms, they can be separated into several categories from the different points of views. For the comparison purposes, we select the following algorithms which are available in WEKA [2].

- Correlation-based Feature Selection (CFS): CFS searches feature subsets according to the degree of redundancy among the features. The evaluator aims to find the subsets of features that are individually highly correlated with the class but have low inter-correlation. The subset evaluators use a numeric measure, such as conditional entropy, to guide the search iteratively and add features that have the highest correlation with the class.

- Statistics-based Feature Selection: Information Gain (IG) and Chi-Square measures are examples in this category. The Information gain measure evaluates features by computing their information gain with respect to the class. The Chi-square measure evaluates features by ranking the chi-square statistic of each feature with respect to the class.

- Instance-based Feature Selection: Relief is an instance-based method that evaluates each feature by its ability to distinguish the neighboring instances. It randomly samples the instances and checks the instances of the same and different classes that are near to each other. An exponential function governs how rapidly the weights degrade with the distance.

- Transformation-based Feature Selection: Principal Components Analysis (PCA), for example, transforms the set of features to the eigenvectors space. Since each eigenvalue gives the variance along its axis, we could use such a special coordinate system that depends on the cloud of points with a certain variance in each direction. All the components could be used as new features but the first few account for most of the variance in the data set.

The details of these functions available in WEKA could be found in [19]. An important step of using a feature selection method is to set up the stopping criterion, which determines when the feature selection algorithm stops and concludes that the subset found at that point is the best feature subset. Some of the stopping criteria adopted by the feature selection methods in the literature are (i) complete search; (ii) a threshold, such as minimum number of features; (iii) subsequent addition, such as in CFS; and (iv) classification error rate [14].

2.2 Classification Algorithms

There are several categories of classifiers. Some of them are trees, functions, Bayesian classifiers, lazy classifiers, rules-based, and meta-learning algorithms [19]. The most popular classification algorithms in data mining voted in [20] are C4.5 (trees), Support Vector Machine (functions), Naive Bayesian (Bayesian), K-Nearest Neighbor (lazy), Apriori (rules), and Adaptive Boosting (meta-learning). Based on this fact, we chose to use the aforementioned classification algorithms in our experiments with the exception of the association rule-based classification, as WEKA does not include an implementation for that classifier. The following are the definitions of those classifiers that are used in our experiments taken from [10]:

- Decision Tree (C4.5).

  C4.5 decision tree learner is a tree structure where each non-leaf node represents a test on a feature, each branch denotes an outcome of the test, and each leaf node represents a class label. The Decision Tree classifier became popular due to the fact that the construction of a decision tree classifier does not require any domain knowledge, and the acquired knowledge in a tree form is easy to understand. In addition, the classification step of decision tree induction is simple and fast. C4.5 uses the information gain ratio as its feature selection measure. Besides the splitting criterion, another interesting challenge of building a decision tree is to overcome the over-fitting of the data. To achieve that, C4.5 uses a method called pessimistic pruning.

- Support Vector Machine (SVM).

  Support Vector Machine is built on the structural risk minimization principle to seek a decision surface that can separate the data points into two classes with a maximal margin between them. The choice of the proper kernel function is the main challenge when using a SVM. It could have different forms such as Radial Basis Function (RBF) kernel and polynomial kernel. The advantage of the SVM is its capability of learning in sparse, high-dimensional spaces with very few training examples by minimizing a bound on the empirical error and the complexity of the classifier at the same time. WEKA uses the Sequential Minimal Optimization (SMO) algorithm for SVM.

- Naive Bayesian (Bayes).

  The Bayesian classifier is a statistical classifier, which has the ability to predict the probability that a given instance belongs to a particular class. The probabilistic Naive Bayes classifier is based on Bayes's rule and assumes that given the class, features are independent, which is called class conditional independence. In theory, Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, this is not always the case in practice, because of the previously mentioned assumption. Even so, the Naive Bayesian classifier has exhibited high accuracy and high speed when applied to large databases.

- K-Nearest Neighbor (KNN).

  The K-Nearest Neighbor algorithm is used under the assumption that instances that are closer to each other generally belong to the same class. Thus KNN is an instance-based learner. The testing sample is labeled according to the class of its first K nearest neighbors. The weight is converted form the distance between the test instance and its predictive neighbors in the training instances. As new training instances are added, the oldest ones are removed to keep the number of training instances at the size of K. The most common metric for computing the distance is the Euclidean distance. For nominal data, the distance between instances according to a particular feature is 0 if their values are the same and 1 otherwise.

- Adaptive Boosting (AdaBoost).

  Boost is a general strategy to improve the accuracy of the classifiers. In boosting, weights are assigned to the training instances and a series of classifiers is iteratively learned. WEKA includes the Adaptive Boosting M1 method. One advantage of the AdaBoost is that it is fast. It can be accelerated by specifying a threshold for weight pruning.

3 The Proposed Framework

The algorithm proposed in this paper achieves the goal of reducing the feature space of a semantic concept detection system by applying Multiple Correspondence Analysis (MCA) to multimedia feature data.

3.1 Multiple Correspondence Analysis (MCA)

Multiple correspondence analysis (MCA) extends the standard Correspondence Analysis (CA) by providing the ability to analyze tables containing some measure of correspondence between the rows and columns with more than two variables [15]. In its basic format, a multimedia database stores features (attributes) and class labels for several instances such as frames, shots, or scenes, for example. If we consider the instances as the rows in the MCA table, and the features (attributes) and class labels as the columns of that table, we can see that when using MCA, the correspondence between the features and the classes could be captured, which could help us narrow the semantic gap between the low level features and the concepts (class labels) in a multimedia database.

MCA is used to analyze a set of observations described by a set of nominal variables, and each nominal variable comprises several levels. In general, the features that are extracted from multimedia streams are numerical in their nature. Therefore, in order to be able to properly use MCA, the extracted quantitative features should be quantized into bins in some manner. Assuming that there are I rows and K columns in the MCA table, the nominal features will have J_k levels, and the total number of items (bins) will be equal to J. Therefore, if there are I data instances in a multimedia database, which are characterized by a set of low-level features, after discretization (i.e., converting the numerical features into nominal ones), there will be K nominal features (including the classes), and each feature will have J_k items (feature-value pairs).

Next, MCA will scan the discretized data to generate the indicator matrix. The indicator matrix is a binary representation of the different categorical values. Each column in this matrix represents a level (item) generated during the data discretization process, while each row represents an instance. The indicator matrix will indicate the appearance of items using the value 1. For a specific instance, only one level (item) can be present for each feature, and therefore each feature can only have one value of 1 in the indicator matrix for each instance. Standard CA analyzes the indicator matrix; while MCA calculates the inner product of the indicator matrix, which generates the Burt matrix $Y = X^T X$, and uses it later for analysis. The size of the indicator matrix is $I \times J$, and the size of the Burt matrix is $J \times J$.

Now, let the grand total of the Burt matrix be N and the probability matrix be $Z = Y/N$. The vector of the column totals of Z is a $1 \times J$ mass matrix M, and $D = diag(M)$. Furthermore, let Δ be the diagonal matrix of the singular values, the columns of P be the left singular vectors (gene coefficient vectors), and the rows of Q^T be the right singular vectors (expression level vectors) in the singular value decomposition (SVD) theorem. MCA will provide the principle components using SVD as follows.

$$D^{-\frac{1}{2}}(Z - MM^T)(D^T)^{-\frac{1}{2}} = P\Delta Q^T. \quad (1)$$

Finally, the multimedia data could be projected into a new space by using the first and the second principle components discovered using Equation 1. The weight of the correlation between the different items and the different classes can be used as an indication to the similarity between them. Such similarity could be calculated as the inner product of each item and each class, i.e., the cosine of the angle between each item and each class. Since the difference between an item and a class ranges from 0 to 180 degrees, and the cosine function decreases from 1 to -1 for that range, the higher correlated item and class would be the pairs that project to the new space with a smaller angel between them.

3.2 MCA-based Feature Selection

The proposed framework consists of several stages as can be seen in Figure 1.

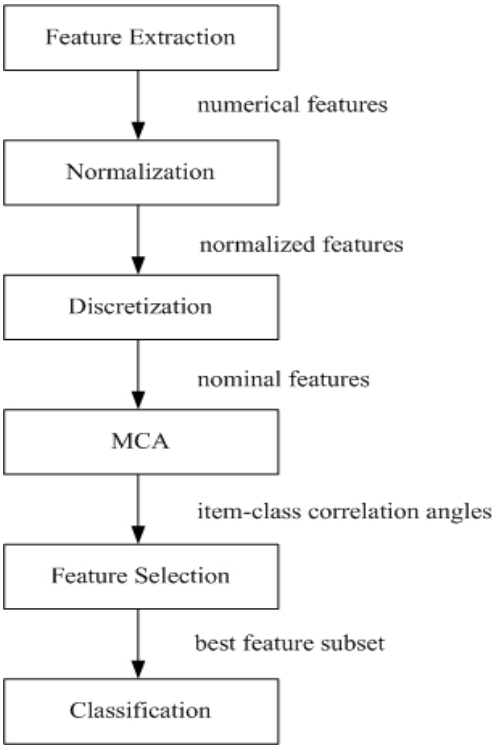

Figure 1. The Proposed Framework.

First, the audiovisual low-level features are extracted from the data. A total of 28 different features, including

11 visual features, 16 audio features, and 1 feature that represents the length of the shot are extracted. The normalization process is done right after the features have been extracted. Next, we discretize the data in order to be able to properly use MCA, since all the features are numerical and MCA requires the input data to be nominal. We discretize the training data set first, and then use the same partitions for discretizing the testing data set. Each interval of the discretization is considered as an item.

Following that, MCA is applied to the discretized training data set and the angles between each item and each class are computed. As mentioned before, the angle between an item and a class has been observed to quantify the correlation between them, and therefore, we have decided to use the angle as our stopping criterion for the proposed feature selection algorithm. One possible threshold condition could be those items whose angle values are smaller than 90 degrees, but it may not be a good choice. In order to determine the proper angle threshold value, the angles generated by MCA for each concept are sorted in the ascending order. We used the first big gap from the distribution of the sorted angles before 90 degree as the lower boundary, and used 90 degrees as the upper boundary. The average of the angles falling into this range was used as our threshold value. Based on this threshold value, the items which have the corresponding angle values that were smaller than the threshold value were kept. This automatic procedure facilitates our proposed framework the capability of identifying different angle thresholds for positive and negative classes. Therefore, we could always get enough best selected items to identify the positive class from the data set.

After the different items generated by the discretization stage are evaluated, the best features will be selected when most of the items from a particular feature are kept. Finally, the classification for the selected semantic concepts by using the aforementioned five well-known classifiers is conducted.

4 Experiments and Results

To evaluate the proposed framework, we used the news broadcast and movies provided by TRECVID [1]. Using 23 video data as our testbed, we have chosen five concepts, namely vegetation, sky, outdoor, crowd, and face. These concepts were taken from the list of concepts provided for the TRECVID 2007 high level feature extraction task. We chose these concepts because (i) there are sufficient amounts of instances to build useful training and testing data sets for these concepts, and (ii) these concepts represent both balanced and imbalanced data sets, which allows us to demonstrate the robustness of our framework. The concept names and their corresponding definitions from [1] are discussed as follows. The ratio of the number of positive instances to the number of negative instances for each concept is listed in Table 1.

- Vegetation: Shots depicting natural or artificial greenery, vegetation woods, etc.;
- Sky - Shots depicting sky;
- Crowd - Shots depicting a crowd;
- Outdoor - Shots of outdoor locations;
- Face: Shots depicting a face.

Concept Name	P / N Ratio
Vegetation	0.12
Sky	0.14
Crowd	0.28
Outdoor	0.51
Face	0.96

Table 1. Positive (P) to Negative (N) instance ratio per investigated concept.

We evaluated our system using the precision (Pre), recall (Rec), and F1-score (F1) metrics under the 3-fold cross validation approach, i.e., three different random sets of training and testing data sets were constructed for each concept. To show the efficiency of our proposed framework, we compared its performance to the performance of the four different feature selection algorithms using five different classifiers available in WEKA [2]. The average precision, recall, and F1-score values of all the feature selection methods for the aforementioned concepts are shown in Table 2 through Table 6. These tables can be read as follows: columns 3 to 7 represent the different feature selection algorithms we have used. These algorithms are: Correlation-based feature selection (CFS), Information gain (IG), Relief (RE), Principal components analysis (PCA), and Multiple correspondence analysis (MCA). The rows represent the different classification algorithms used as follows: Decision tree (DT), Support vector machine (SVM), Naive Bayesian (NB), K-nearest neighbor (KNN), and Adaptive boosting (AB).

We observe that SVM always yields zero precision and recall in those concepts with extremely imbalanced data, namely vegetation (ratio=0.12) and sky (ratio=0.14). This is because when the class distribution is too skewed, SVM will generate a trivial model by predicting everything to the majority class, i.e., the negative class. In the case of the Information Gain, Relief, and PCA methods, WEKA produced a ranked list of the features without performing the actual feature selection. Due to this fact, we had to select the best stopping criteria. After an extensive empirical study, we have set the stopping criteria for these three

		CFS	IG	RE	PCA	MCA
DT	Pre	0.00	0.00	0.00	0.00	0.17
	Rec	0.00	0.00	0.00	0.00	0.01
	F1	0.00	0.00	0.00	0.00	0.02
SVM	Pre	0.00	0.00	0.00	0.00	0.00
	Rec	0.00	0.00	0.00	0.00	0.00
	F1	0.00	0.00	0.00	0.00	0.00
NB	Pre	0.00	0.35	0.28	0.13	0.38
	Rec	0.00	0.03	0.02	0.01	0.09
	F1	0.00	0.05	0.04	0.01	0.14
KNN	Pre	0.00	0.50	0.35	0.19	0.37
	Rec	0.00	0.05	0.04	0.06	0.11
	F1	0.00	0.09	0.06	0.09	0.16
AB	Pre	0.00	0.00	0.00	0.00	0.33
	Rec	0.00	0.00	0.00	0.00	0.01
	F1	0.00	0.00	0.00	0.00	0.01

Table 2. Average precision (Pre), recall (Rec) and F1-score (F1) for "vegetation" over five classifiers

methods as follows. We have calculated the average score of each of the previously mentioned ranked lists and used this average value as a threshold for selecting the features, i.e., those features that had a higher score than the average value were selected as the best subset of features produced by these three methods.

As can be observed in Tables 2 through 6, our proposed framework achieves promising results compared to all the other feature selection methods over all classifiers, especially in the cases of the imbalanced data sets (i.e., vegetation, sky, and crowd concepts). We can further observe that the recall values and the F1-scores for the proposed framework are always higher over all the classifiers. This encouraging observation demonstrates the fact that the proposed framework has the ability to help the classifiers to detect more positive instances in the testing data set without misclassifying too many negative instances by identifying the best feature subset for each of the investigated concepts. In addition, the proposed framework was able to reduce the feature space by approximately 50% for all the investigated concepts in the experiments, which is considered a significant feature space reduction. This demonstrates that the proposed framework can better represent the semantic concepts using the reduced feature set.

5 Conclusions

In this paper, a correlation-based and transformation-based feature selection framework using MCA is proposed to handle multimedia semantic understanding related problems such as high-dimensionality, semantic gap, and the

		CFS	IG	RE	PCA	MCA
DT	Pre	0.59	0.42	0.17	0.78	0.59
	Rec	0.06	0.04	0.03	0.05	0.08
	F1	0.11	0.07	0.05	0.09	0.14
SVM	Pre	0.00	0.00	0.00	0.00	0.00
	Rec	0.00	0.00	0.00	0.00	0.00
	F1	0.00	0.00	0.00	0.00	0.00
NB	Pre	0.44	0.23	0.32	0.47	0.37
	Rec	0.20	0.31	0.13	0.10	0.47
	F1	0.28	0.26	0.17	0.16	0.40
KNN	Pre	0.51	0.50	0.41	0.36	0.47
	Rec	0.12	0.13	0.12	0.20	0.19
	F1	0.19	0.20	0.17	0.25	0.27
AB	Pre	0.60	0.38	0.35	0.62	0.59
	Rec	0.02	0.03	0.02	0.04	0.06
	F1	0.04	0.06	0.04	0.08	0.12

Table 3. Average precision (Pre), recall (Rec) and F1-score (F1) for "sky" over five classifiers

		CFS	IG	RE	PCA	MCA
DT	Pre	0.60	0.80	0.62	0.37	0.57
	Rec	0.19	0.10	0.19	0.16	0.23
	F1	0.28	0.15	0.28	0.22	0.32
SVM	Pre	0.82	0.79	0.79	0.33	0.79
	Rec	0.06	0.08	0.08	0.01	0.08
	F1	0.11	0.15	0.15	0.02	0.15
NB	Pre	0.49	0.49	0.48	0.48	0.46
	Rec	0.40	0.31	0.34	0.17	0.41
	F1	0.44	0.37	0.39	0.25	0.44
KNN	Pre	0.46	0.60	0.55	0.43	0.48
	Rec	0.29	0.20	0.23	0.35	0.31
	F1	0.36	0.29	0.30	0.37	0.38
AB	Pre	0.63	0.76	0.77	0.66	0.64
	Rec	0.12	0.10	0.04	0.05	0.14
	F1	0.19	0.16	0.07	0.09	0.22

Table 4. Average precision (Pre), recall (Rec) and F1-score (F1) for "crowd" over five classifiers

		CFS	IG	RE	PCA	MCA
DT	Pre	0.59	0.59	0.58	0.54	0.60
	Rec	0.47	0.41	0.41	0.44	0.48
	F1	0.52	0.48	0.48	0.49	0.53
SVM	Pre	0.58	0.57	0.38	0.59	0.64
	Rec	0.34	0.32	0.22	0.33	0.42
	F1	0.43	0.41	0.28	0.42	0.50
NB	Pre	0.54	0.53	0.52	0.58	0.53
	Rec	0.51	0.51	0.44	0.39	0.54
	F1	0.53	0.52	0.46	0.47	0.54
KNN	Pre	0.58	0.56	0.53	0.50	0.54
	Rec	0.50	0.45	0.49	0.50	0.57
	F1	0.53	0.50	0.50	0.50	0.56
AB	Pre	0.59	0.59	0.56	0.56	0.58
	Rec	0.32	0.35	0.21	0.29	0.39
	F1	0.41	0.43	0.30	0.38	0.47

Table 5. Average precision (Pre), recall (Rec) and F1-score (F1) for "outdoor" over five classifiers

		CFS	IG	RE	PCA	MCA
DT	Pre	0.65	0.65	0.66	0.68	0.65
	Rec	0.61	0.62	0.58	0.54	0.63
	F1	0.63	0.63	0.62	0.61	0.64
SVM	Pre	0.67	0.67	0.66	0.66	0.68
	Rec	0.63	0.63	0.63	0.62	0.64
	F1	0.65	0.65	0.64	0.64	0.66
NB	Pre	0.64	0.64	0.64	0.65	0.65
	Rec	0.65	0.65	0.64	0.59	0.66
	F1	0.65	0.65	0.64	0.60	0.65
KNN	Pre	0.61	0.61	0.62	0.62	0.62
	Rec	0.71	0.67	0.67	0.65	0.73
	F1	0.66	0.64	0.64	0.64	0.67
AB	Pre	0.64	0.64	0.64	0.66	0.63
	Rec	0.63	0.65	0.64	0.54	0.68
	F1	0.64	0.64	0.64	0.69	0.65

Table 6. Average precision (Pre), recall (Rec) and F1-score (F1) for "face" over five classifiers

imbalance data in a multimedia database. The TRECVID 2007 benchmark data is used to evaluate the concept detection performance of our proposed framework compared with several widely used feature selection schemes under several well-known classifiers. We utilize the functionality of MCA to measure the correlation between extracted low-level audiovisual features and classes to infer the high-level concepts (semantics). The experimental results show that our proposed framework performs well in improving the detection of the high-level concepts, namely vegetation, outdoor, sky, crowd, and face. Furthermore, the results demonstrate the superiority of the proposed framework over the other feature selection methods used in the case of the imbalanced data over all five classifiers. The proposed feature selection framework proves to play a major role in assisting semantic concept detection systems to better understand the semantic meaning of multimedia data under real world constraints such as imbalanced data sets.

6 Acknowledgement

For Shu-Ching Chen, this research was supported in part by NSF HRD-0317692 and Florida Hurricane Alliance Research Program sponsored by the National Oceanic and Atmospheric Administration.

References

[1] *Guidelines for the TRECVID 2007 Evaluation*. http://www-nlpir.nist.gov/projects/tv2007/tv2007.html.

[2] *WEKA*. http://www.cs.waikato.ac.nz/ml/weka/.

[3] A. Amir, S. Basu, G. Iyengar, C.-Y. Lin, M. Naphade, J. R. Smith, S. Srinivasan, and B. Tseng. A multimodal system for the retrieval of semantic video events. *Computer Vision and Image Understanding Archive*, 56(2):216–236, November 2004.

[4] 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, Special Issue on Semantic Retrieval of Multimedia*, 23(2):38–46, March 2006.

[5] S.-C. Chen, N. Zhao, and M.-L. Shyu. Modeling semantic concepts and user preferences in content-based video retrieval. *International Journal of Semantic Computing*, 1(3):377–402, September 2007.

[6] S. Das. Filters, wrappers and a boosting-based learning. *Proc. International Conference of Machine Learning*, pages 74–81, 2001.

[7] J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. *Proc. International Conference of Machine Learning*, pages 247–254, 2000.

[8] A. Estabrooks, T. Jo, and N. Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36, February 2004.

[9] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. *Proc. International Conference of Machine Learning*, pages 359–366, 2000.

[10] J. Han and M. Kamber. *Data Mining: Concepts and Techniques, 2nd Edition*. The Morgan Kaufmann, 2006.

[11] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

[12] Y. Kim, W. Street, and F. Menczer. Feature selection for unsupervised learning via evolutionary search. *ACM SIGKDD International Conference of Knowledge Discovery and Data Mining*, pages 365–369, 2000.

[13] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Video semantic concept discovery using multimodal-based association classification. *IEEE International Conference on Multimedia and Expo*, pages 859–862, July 2007.

[14] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. on Knowledge and Data Engineering*, 17(4):491–502, April 2005.

[15] N. J. E. Salkind. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage Publications, Inc, 2007.

[16] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56(3):151–177, February 2004.

[17] M.-L. Shyu, S.-C. Chen, Q. Sun, and H. Yu. Overview and future trends of multimedia research of content access and distribution. *International Journal of Semantic Computing*, 1(1):29–66, March 2007.

[18] G. M. Weiss and F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, October 2003.

[19] I. H. Witten and E. Frank. *Data Mining Practical Machine Learning Tools and Techniques, 2nd Edition*. Morgan Kaufmann, 2005.

[20] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Y. Q., H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, pages 1–37, December 2007.

[21] E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. *Proc. International Conference of Machine Learning*, pages 601–608, 2001.

[22] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. *Proc. International Conference of Machine Learning*, pages 856–863, 2003.