CONCEPT LEARNING FOR IMAGE AND VIDEO RETRIEVAL: THE INVERSE RANDOM UNDER SAMPLING APPROACH

M. A. Tahir, J. Kittler, F. Yan and K. Mikolajczyk

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, GU2 7XH, UK
E-mail: {m.tahir,j.kittler,f.yan,k.mikolajczyk}@surrey.ac.uk

ABSTRACT

A typical concept-detection problem is characterised by greatly disproportionate sizes of the populations of training samples in the concept and anti-concept classes. In many cases, the population of anti-concept (negative) examples outnumber the concept examples. In this paper, an inverse random under sampling method is proposed to solve this imbalance problem. By the proposed method of inverse under sampling of the anti-concept class we can construct a large number of concept detectors which in the fusion stage facilitate a fine control of both false negative rates and false positive rates. In this method the main emphasis in learning the discriminant functions is on the concept class, leading to an almost perfect separation of the two classes for each detector. The proposed methodology is applied to commonly-used video and image collection benchmarks: Mediamill and Scene datasets. The results indicate significant performance gains. For some concepts, the improvement in the average precision is by several orders of magnitude, and the mean average precision is 12% and 17% better for Mediamill and Scene datasets respectively when compared with conventionally trained logistic regression classifier.

1. INTRODUCTION

The image/video database retrieval problem involves finding in the database, instances of multimedia content that is similar to the content of interest, specified by the user. The required content can be specified by example, or it can be defined abstractly in terms of concepts. In the former case we refer to the retrieval problem as *content based retrieval*, and in the latter case as *concept-detection*. In this paper we shall focus on the latter variant where it is assumed that for each concept we have a set of representative examples (images). A machine learning algorithm is then used to construct a model of the concept class that can successfully discriminate concept samples from negative (anti-concept) instances.

Mathematically, the concept-detection problem can be formulated as either one class or conventional two-class pattern recognition problem. In the former case we build a generative model of the concept class and the concept-detection then involves testing the hypothesis that an unknown image is consistent with the model, i.e. could have been generated by it. The alternative is to view the concept retrieval as a two-class problem where the second class is represented by negative samples, i.e. images that do not contain the specified concept. The problem can be solved using generative or discriminative models learnt using the training data.

In many practical situations the number of examples representing the concept class is very limited. This precludes building a reliable generative model and, in consequence, the approaches which rely on such models are inappropriate. Thus most of the image/video database methods in the literature adopt the two-class formulation and, as the rest of this paper, resort to discriminative machine learning solutions. For example, in [3], Support Vector Machines (SVMs) are used in a hierarchical manner for image annotation and retrieval. In [15], the goal is to detect the presence of 101 semantic concepts in videos. The detection of each concept is formulated as a binary classification problem. In this discriminative setting, again SVMs are employed. [8] considers the situation where each image/video can take multiple class labels, i.e., the multi-label problem. However, the underlying probabilistic model is still discriminative. More examples of using discriminative machine learning techniques for image/video classification can be found in [5, 7, 9, 13].

A typical concept-detection problem is characterised by greatly disproportionate sizes of the populations of training samples in the concept and anti-concept classes. For the negative class it is very easy to compile a large training set, which is invariable constituted by fusing the training samples of all the other concept classes to form the negative class. With this approach, the relative sizes of the concept class and negative sample class training sets can differ by several orders of magnitude. This huge disparity in the training set cardinalities poses a challenging machine learning problem when designing the concept detectors.

The problem of disproportionate class size is not unique to image/video retrieval. In has been encountered in other applications in pattern recognition and statistics. A number of different sampling strategies have been suggested to deal with it. The possibilities explored in the literature include stratified sampling [2] where the same number of training samples is drawn for each class. As the negative class in stratified sampling becomes under sampled, this approach opens the possibility of drawing a large number of different anti-concept training sets and designing multiple classifiers that can then be fused to improve the detection performance.

Classifier designs based on disproportionate training sets sizes manifest themselves in exhibit conditional classification errors that are dependent on the population size probabilities. We shall argue that in order to achieve good retrieval performance, as measured in terms of *average precision*, it is essential to provide the designer with a very fine control over both false positive rate (incorrectly detecting negative samples as belonging to the concept class) and false negative rate.

In this paper, a novel inverse random under sampling (IRUS) method is proposed for the class imbalance problem in which the ratio of the respective training set cardinalities is inversed. The idea is to severely under sample the negative class (majority class), thus creating a large number of distinct negative training sets. For each training set we then find a linear discriminant which separates the positive class from the negative samples. As the number of positive samples in each training set is greater than the number of negative samples, the focus in machine learning is on the positive class and consequently it can invariably be successfully separated from the negative training samples. Thus each training set yields one classifier design. By combining the multiple designs, we construct a composite between the positive class and the negative class. We shall argue that this boundary has the capacity to delineate the positive class more effectively than the solutions obtained by conventional learning.

The proposed methodology is applied to an image database and a video database involving 6 and 39 concepts respectively. We use standard benchmarking sets, namely the Mediamill Challenge video database [15] and Scene database [17, 3], for which the state of the art performances are well documented in the literature and features are pre-computed and available on line. We validate the advocated approach and demonstrate that it yields significant performance gains. In the case of some concepts the improvement in the average precision is by several orders of magnitude, and the mean average precision is 12% and 17% better for Mediamill and Scene datasets respectively.

The paper is organised as follows. Section 2 provides briefly review several class imbalance methods followed by proposed inverse random under sampling method (IRUS) in section 3. Section 4 describes the experimental setup followed by results and discussion in Section 5. The paper is drawn to conclusion in Section 6.

2. RELATED WORK

The most commonly used methods to handle imbalanced data sets involve under sampling or over sampling of the original data sets. Random over sampling and random under sampling are the most popular non-heuristic methods that balance class representation through random replication of the minority class and random elimination of majority class examples respectively. There are some limitations of both random under sampling and random over sampling. For instance, under-sampling can discard potentially useful data while over-sampling can increase the likelihood of overfitting [1]. Despite these limitations, random over sampling in general is among the most popular sampling techniques and provides competitive results when compared with most complex methods [1, 11].

Several heuristic methods are proposed to overcome these limitations including Tomek links, Condensed Nearest Neighbour Rule (CNN), One-sided selection and Neighbourhood Cleaning rule (NCL) are several well-known methods for under-sampling [10] while Synthetic Minority Over-Sampling Technique (SMOTE) is a well-known method for over-sampling technique [6]. The main idea in SMOTE is to generate synthetic examples by operating in the “feature space” rather than the “data space” [6]. The minority class is oversampled by interpolating between several minority class examples that lie together. Depending upon the amount of over-sampling required, neighbours from the k nearest neighbours are randomly chosen. Thus, the overfitting problem is avoided and the decision boundaries for the minority class are spread further into the majority class space [1].

Liu et al [11] and Chan et al [4] examine the class imbalance problem by combining classifiers built from multiple undersampled training sets. In both approaches, several subsets from the majority class with each subset having approximately the same number of samples as the minority class are created. One classifier is trained from each of these subsets and the minority class and then the classifiers are combined. Both these approaches differ in grouping multiple classifiers and in creating subsets from the majority class.

3. PROPOSED INVERSE RANDOM UNDER SAMPLING METHOD (IRUS)

In this section, we will discuss the proposed inverse random under sampling (IRUS) method. For convenience, we refer to the minority class as the concept class and the majority class as the anti-concept class. A conventional training of a concept detector using a data set containing representative proportions of samples from the concept and anti-concept classes will tend to find a solution that will be biased towards the larger class. In other words, the probability of misclassifying samples from the anti-concept class will be lower than the probability of error for the concept class. However, the actual performance will be determined by the underlying overlap of the two-classes and the class prior probabilities. Thus, we need to control the probability of misclassification of samples from the anti-concept class to achieve the target performance objectives. This may require setting the operating point of the detector so as to achieve false positive rate that is lower than what would be yielded by conventional training. This could be achieved by biasing the decision boundary in favour of the anti-concept sample error rates using threshold (off set) manipulation. Alternatively, we could increase the imbalance between the number of samples from the two-classes artificially by eliminating some of them. The latter solution is not very sensible, as we would be depleting the class which is naturally underrepresented even further. The former solution would lead to a substantial increase in the false negative rate.

The problem of learning decision functions in situations involving highly imbalanced class sizes is sometimes mitigated by stratified sampling. This aims to create a training set containing a comparable numbers of samples from all the classes. Clearly, in stratified sampling the training set size would be determined by the number of samples in the underrepresented class. This would lead to a drastic subsampling of the anti-concept class with the resultant reduction in the accuracy of the estimated class boundary. This loss of accuracy can be recovered by means of multiple classifier methodology. By drawing randomly multiple subsets from the anti-concept class data set, each adhering to the stratified sampling criteria, we can design several detectors and fuse their opinions. For a typical imbalance of priors of say 100 : 1, the number of the designs would be too low to allow an alternative approach to controlling false positive error rate and one would have to resort to the biasing methods discussed earlier.

Suppose we take the data set manipulation to the extreme and inverse the imbalance between the two-classes. Effectively we would have to draw sample sets from the anti-concept class of size proportional to P^2 where P is the prior probability of the concept class. This would lead to very small sample sets for the anti-concept class and therefore, a poor definition of the boundary between the two-classes. Nevertheless, the boundary would favour the concept class. Also, as the number of samples from the negative class is very small in relation to the dimensionality of the feature space, the capacity of each boundary to separate the classes fully is high. Moreover, as the number of samples drawn is proportional to P^2, the number of independent sets that can be drawn will be of the order of $\frac{1}{P^2}$. This large number of designs could then be used for controlling the false positive rate using a completely different mechanism. This contrasts with the complex task of biasing a decision boundary in high dimensional space.

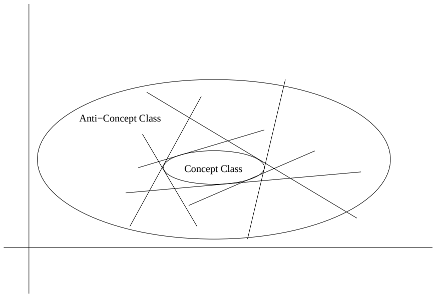

Figure 1: Schematic diagram showing each boundary partitions the training data set by a hyperplane tangent to the surface of the volume occupied by the concept class.

Interestingly, there is another important benefit of the the IRUS method. As the number of samples forming the negative class is very small, each detector design will be significantly different. This will produce highly diverse detectors which are required for effective classifier fusion. The fused decision rule achieves better class separation than a single boundary, albeit estimated using more samples. This is conveyed schematically in Figure 1. Each boundary partitions the training data set by a hyperplane tangent to the surface of the volume occupied by the concept class. It is the union of these tangent hyperplanes created by fusion, which constitutes a complex boundary to the concept class. Such boundary could not easily be found by a single linear discriminant function. If one resorted to nonlinear functions, the small sample set training would most likely lead to a over fitting and, consequently, to poor generalisation on the test set. Figure 2 provides supporting evidence for the above conjecture. The histogram of discriminant function values (i.e. distance from the decision boundary) generated by one thousand classifiers designed using the inverse imbalance sampling principle for a single negative class test sample (blue bar) shows many of the classifiers scoring positive values which lie on the concept class side of the boundary. This is expected for more than half of the classifiers, as the negative sample will lie beyond the concept class, but nevertheless on the same side as the concept class. In contrast, discriminant function values for a single positive class test sample show that most of the classifiers scoring positive values lie on the concept side of the boundary.

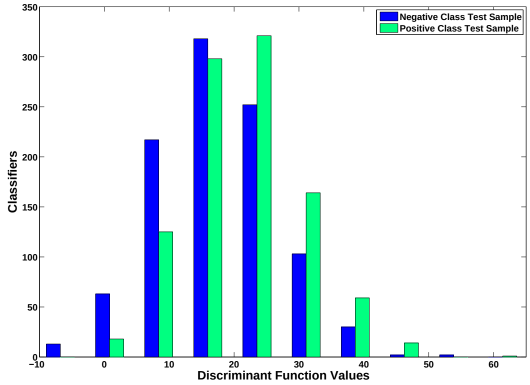

Figure 2: Histogram of Discriminant Function generated by one thousand classifiers.

In summary, we propose a classifier design approach which is based on an inverse imbalance sampling strategy. This is accomplished by combining the outputs of the multiple concept detectors in the fusion stage and thus allows a very accurate definition of the boundary between the concept class and the anti-concept class.

The pseudo code of IRUS is shown in Algorithm 1. S and $Sets$ are user specified parameters. S controls the number of negative samples drawn at random in each model with values ranging from 1 to $N_c - 1$. $Sets$ determine the number of models or classifiers with values greater than N_a/S. For each set Ξ'_a paired with Ξ_c we learn a model h_i. For each model h_i, the probability of unseen instances belonging to concept class D_c is calculated. The probabilities from all models are added. The output is a probability set Ξ_p of the test instances belonging to concept class. Ξ_p is then used to calculate the performance measure discussed in Section 4.3.

Algorithm 1 PseudoCode for Inverse Random Under Sampling (IRUS)

```
Require: $\Xi_c$: Training set of concept patterns with cardinality $N_c$
  $\Xi_a$: Training set of anti-concept patterns with cardinality $N_a$
  $\Xi_t$: Test set with cardinality $N_t$
  $S$: Number of samples from $\Xi_a$ for each Model
  $Sets$: Number of classifiers
Ensure: $\Xi_p$: Probability set of Test instances belonging to concept
  class
  $\Xi_p \Leftarrow 0$
  for $i = 1$ to $Sets$ do
    $\Xi'_a \Leftarrow$ Randomly pick $S$ samples without replacement from $\Xi_a$
    $T_s \Leftarrow \Xi'_a + \Xi_c$
    Train base classifier $h_i$ using $T_s$ samples
    for $j = 1$ to $N_t$ do
      $D_c \Leftarrow$ Probability distribution of Test Sample $\Xi_{tj}$ belonging
      to concept class from $h_i$
      $\Xi_{pj} \Leftarrow \Xi_{pj} + D_c$
    end for
  end for
```

4. EXPERIMENT DESIGN

4.1 Datasets

The effectiveness of the proposed classifier is tested on Mediamill challenge [15] and Scene [3] benchmarks. The mediamill challenge by Snoek et al [15] provides an annotated video dataset, based on the training set of NIST TRECVID 2005 benchmark [14]. This dataset consists of 86 hours of video, divided into a training set (70% of the data or 30993 examples) and test set (30% the data or 12914 examples). On this dataset, the 39 LSCOM-Lite categories are used [12, 16]. The feature vector used in these experiments consists of 120 visual features and available on-line [1] (Experiment1 in mediamill challenge by Snoek et al [15]). In scene dataset [3, 17], 6 categories are used. This dataset is divided into 1211 training samples and 1196 test samples. The feature vector consists of 294 features and available on-line [2]. Table 1 and 2 show the ground truths in both Mediamill and Scene datasets. Detail description about how these feature are computed can be found in [3, 15].

4.2 Benchmark Methods

Logistic Regression (LR) is used as the base classifier for the proposed inverse random under sampling technique (IRUS). The IRUS method is compared with the baseline performance based on the SVM classifier with *RBF* kernel [15] and Hybrid Ensemble Boosting Learning method (HMLB) reported in [17]. In addition, we have compared IRUS with sampling techniques Random Under Sampling (RUS), Random Over Sampling (ROS) and SMOTE. For all sampling techniques, LR is used as a base classifier. We have also compared IRUS with ensemble techniques Bagging and AdaBoost with decision tree (C4.5) as base classifier. The WEKA [18] implementation is used for LR, SMOTE, Bagging and AdaBoost.

4.3 Evaluation Measure

Average precision is standard image ranking measure and is used in this paper. The average precision is a single-valued measure that is proportional to the area under a precision-recall curve. This value is the average of the precision over all relevant judged shots. This metric combines precision and recall into one performance value. This measure is computed from the ranking list of all the key frames in the database established by ordering their similarities to a specified concept. Average Precision for each concept (AP) is defined as

$$AP = \frac{1}{|R|} \sum_{k=1}^{|R|} c_k \qquad (1)$$

where R is the complete set of the positive samples in a test set and the contribution c_k of the k^{th} element in the ranking list is defined as

$$c_k = \begin{cases} \frac{|R \cap M_k|}{k} & if \quad concepttrue \\ 0 & if \quad conceptnottrue \end{cases} \qquad (2)$$

where $M_k = \{i_1, i_2, ..., i_k\}$ is a ranked list of the top k retrieved samples from the test set.

5. RESULTS AND DISCUSSION

5.1 Experiment1: Video Benchmark

Table 1 shows the average precision (AP) for each concept using various methods for the mediamill challenge. From the results, it is observed that IRUS method has highest performance in 15 out of 39 concepts. For concepts like animal, court, natural-disaster, police-security, prisoner, screen, snow, and waterbody, improvement of over 50% is achieved over LR and SVM. However for some concepts, especially flag-usa and sky, there is decrease in performance. This is explained by the fact that for these concepts, some anti-concept samples are ranked very high; i.e. these samples always lie

[1] http://www.science.uva.nl/research/mediamill/challenge/

[2] http://mlkd.csd.auth.gr/multilabel.html

Concept	Ground Truth		LR	SVM [15]	HMLB [17]	RUS	ROS	SMOTE	Bagg	AdaBoost	IRUS
	Train (N_c)	Test		*RBF*							$S = N_c^{0.3}$
Court	63	39	0.069	0.093	0.000	0.004	0.026	0.048	0.047	0.061	**0.189**
Prisoner	103	28	0.005	0.047	0.000	0.003	0.006	0.005	0.007	0.048	**0.137**
Snow	126	68	0.054	0.085	0.108	0.011	0.024	0.042	0.152	0.131	**0.170**
Bus	132	83	0.011	**0.013**	0.000	0.008	0.010	0.011	0.009	**0.013**	0.011
Explosion	164	134	0.081	**0.098**	0.076	0.025	0.056	0.079	0.077	0.049	0.083
Charts	234	66	0.290	0.327	0.171	0.017	0.106	0.224	**0.347**	0.300	0.282
Boat	242	70	0.043	0.096	0.121	0.044	0.031	0.041	0.112	**0.183**	0.138
Desert	250	186	0.106	0.103	**0.174**	0.065	0.094	0.113	0.110	0.080	0.142
Natural Disaster	250	120	0.048	0.055	0.065	0.027	0.042	0.046	0.059	0.057	**0.113**
Flag USA	285	121	0.183	**0.227**	0.171	0.043	0.145	0.212	0.130	0.085	0.087
Police/Security	286	100	0.018	0.012	0.000	0.013	0.017	0.017	0.054	0.029	**0.100**
Aircraft	306	122	0.080	0.073	**0.187**	0.046	0.066	0.082	0.086	0.074	0.138
Weather Report	307	161	0.342	0.405	0.307	0.114	0.178	0.319	0.263	0.261	**0.429**
Animal	309	117	0.148	0.209	0.141	0.028	0.062	0.119	0.310	0.395	**0.436**
Maps	358	156	0.378	0.476	0.389	0.110	0.256	0.377	0.386	0.356	**0.504**
Truck	361	132	0.039	0.038	0.040	0.033	0.036	0.036	0.031	0.022	**0.043**
Screen	475	245	0.095	0.101	0.061	0.055	0.073	0.087	0.107	0.095	**0.185**
Office	485	226	0.076	0.077	**0.282**	0.056	0.065	0.079	0.077	0.064	0.108
Mountain	508	131	0.190	0.141	0.134	0.121	0.169	0.181	0.215	0.176	**0.252**
People Marching	597	533	0.261	0.228	**0.332**	0.208	0.245	0.297	0.178	0.165	0.205
Water Body	716	244	0.173	0.150	0.146	0.121	0.150	0.188	0.257	0.272	**0.333**
Corporate-Leader	797	168	0.018	0.016	0.000	0.016	0.018	**0.019**	0.016	0.015	**0.019**
Sports	1166	337	0.211	**0.304**	0.115	0.119	0.139	0.147	0.181	0.184	0.199
Vegetation	1198	599	**0.215**	0.183	0.191	0.193	0.192	0.197	0.161	0.131	0.179
Military	1283	850	**0.242**	0.217	0.250	0.226	0.236	0.241	0.182	0.152	0.239
Meeting	1405	627	0.245	0.257	**0.272**	0.223	0.233	0.234	0.198	0.163	0.233
Car	1509	766	0.232	0.252	**0.253**	0.188	0.208	0.222	0.250	0.233	0.241
Building	2126	1441	0.303	0.316	**0.335**	0.257	0.286	0.293	0.278	0.232	0.297
Road	2404	852	0.190	0.195	0.212	0.177	0.183	0.185	0.196	0.184	**0.198**
Government Leader	2899	1016	**0.235**	0.213	0.202	0.211	0.224	0.224	0.167	0.152	0.217
Sky	3339	1469	**0.535**	0.478	0.373	0.513	0.520	0.511	0.446	0.394	0.451
Crowd	3559	2082	**0.519**	0.480	0.397	0.505	0.513	0.517	0.454	0.414	0.454
Urban	3651	1136	0.217	0.222	0.197	0.204	0.211	0.205	**0.242**	0.215	0.223
Walking/Running	4219	2174	**0.370**	0.353	0.311	0.359	0.363	0.367	0.330	0.286	0.334
Studio	4234	1834	0.640	0.636	0.463	0.606	0.612	0.614	0.660	0.628	**0.666**
Entertainment	6088	1621	0.281	0.166	0.194	0.273	0.277	0.262	**0.323**	**0.293**	0.293
Outdoor	10130	4950	**0.739**	0.688	0.688	0.735	0.736	0.732	0.736	0.703	0.710
Face	19883	8055	0.897	0.895	0.712	**0.898**	0.897	0.895	0.881	0.876	0.895
People	24071	9798	0.941	0.831	0.830	**0.942**	0.941	0.941	0.930	0.918	0.937
MAP ($N_c < 1000$))			0.123	0.140	0.132	0.053	0.085	0.119	0.138	0.133	**0.187**
MAP ($N_c > 1000$))			**0.412**	0.393	0.353	0.390	0.398	0.399	0.389	0.362	0.398
Overall MAP			0.249	0.250	0.228	0.200	0.222	0.241	0.247	0.233	**0.279**

Table 1: Comparison of Average Precision with different methods for Mediamill data set. S = Number of negative samples used for each classifier in IRUS (See Algorithm 1). The second and third columns show the Ground Truths with N_c is the number of samples in the concept class.

Concept	Ground Truth (N_c)		LR	SVM	HMLB [17]	RUS	ROS	SMOTE	Bagg	AdaBoost	IRUS
	Train	Test		*RBF*							$S = N_c^{0.2}$
Mountain	196	235	0.829	**0.896**	0.526	0.812	0.807	0.839	0.884	0.866	0.888
Field	197	200	0.681	0.873	0.533	0.707	0.646	0.673	0.876	**0.899**	0.877
Fall Foliage	199	165	0.922	**0.960**	0.651	0.930	0.924	0.927	0.920	0.948	0.880
Urban	207	204	0.501	0.689	0.371	0.446	0.476	0.478	0.674	0.668	**0.765**
Beach	227	200	0.702	0.772	0.374	0.663	0.680	0.707	0.759	0.727	**0.817**
Sunset	277	256	0.488	0.562	0.349	0.437	0.491	0.509	0.566	0.586	**0.610**
MAP			0.687	0.792	0.467	0.666	0.671	0.689	0.780	0.782	**0.806**

Table 2: Comparison of Average Precision with different methods for Scene data set.

on the positive side of the boundary. Overall, IRUS yields significant performance gains. In the case of some concepts the improvement in the average precision is by several orders of magnitude, and the mean average precision is 12%, 11%, 22%, 39%, 26%, 16%, 13% and 20% when compared with LR, SVM, HMLB, RUS, ROS, SMOTE, Bagging and AdaBoost respectively.

To show the effectiveness of IRUS especially for highly unbalanced concepts, we have divided all concepts in two clusters. One cluster consists of concepts with less than 1000 training samples and other cluster with greater than 1000 training samples. As clear from Table 1, there is improvement of 52%, 34%, 41%, 251%, 119%, 57%, 40%, 49% when compared with LR, SVM, HMLB, RUS, ROS, SMOTE, Bagging and AdaBoost respectively for highly unbalanced concepts from the first cluster. For the second cluster, the performance is almost similar in all methods with LR has highest MAP. Since, the detection of each concept is formulated as a binary classification problem, it is always possible to use separate classifiers for different concepts. Thus, for concepts from second cluster, conventional learning like LR, SVM can easily be adopted.

5.2 Experiment 2: Image Benchmark

Table 2 shows the average precision (AP) for each concept using various methods for the scene dataset. From the results, it is cleared that the proposed method is significantly better in all concepts when compared with LR, HMLB, RUS, ROS and SMOTE. This table also indicates that when IRUS is compared with SVM, Bagging and AdaBoost, performance varies for different object categories. For example, SVM has higher performance in Mountain, AdaBoost performs better in Field and Fall Foliage while IRUS performs higher in Urban, Beach and Sunset. Overall, the mean average precision (MAP) for IRUS is 17%, 2%, 72%, 21%, 20%, 17%, 3%, and 3% when compared with LR, SVM, HMLB, RUS, ROS, SMOTE, Bagging and AdaBoost respectively.

5.3 Discussion

The proposed inverse random subsampling method is very effective for image and video retrieval problems with highly unbalanced data sets. The results clearly indicate that traditional sampling techniques are not effective when dealing with highly unbalanced data sets in the retrieval problems. When IRUS is compared with other sampling methods RUS, ROS and SMOTE improvement of 251%, 119%, and 57% respectively is achieved for highly unbalanced concepts (Cluster1 from Mediamill Challenge). The evaluation of run time parameters (S and $Sets$ See Algorithm 1) are not shown due to lack of space. For mediamill challenge, $S = N_c^{0.3}$ is used for all concepts while for scene $S = N_c^{0.2}$ is used. The other run time parameter, $Sets$ which determine the number of classifiers is equal to $1.5 \times N_a/S$ for both data sets. In this paper, fusion is performed through that the sum of scores obtained from individual classifiers. It is part of our future research to investigate other ways of combination such as Fuzzy Integral, Dempster-Shafer etc to improve the performance.

6. CONCLUSION

A novel inverse random under sampling (IRUS) method is proposed in this paper to solve the imbalance problem in concept-detection. By the proposed method we can construct a large number of concept detectors which in the fusion stage facilitate a fine control of both precision and recall. The main idea is to emphasise learning the discriminant functions for the concept class, leading to almost perfect separation of the two-classes for each detector. The distinctiveness of IRUS is assessed experimentally using image and video benchmarks. The results indicate significant performance gains. For some concepts, the improvement in the average precision is by several orders of magnitude, and the mean average precision is 12% and 17% better for Mediamill and Scene datasets respectively when compared to conventionally trained logistic regression classifiers. In this paper, logistic regression is used as a base classifier. It would be interesting to see how other well-known classifiers like NaiveBayes, SVM, KNN, LDA behave when used as a base classifier in our proposed inverse under sampling method.

Acknowledgement: This work was supported by the European Commission EU VIDI-Video IST-2-045547.

REFERENCES

[1] G. Batista and R. C. Prati amd M. C. Monard. A study of the bahavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6(20–29), 2004.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] M. Boutell, L. Luo amd X. Shen, and J. Luo. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[4] P. K. Chan and S. J.Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 164–168, New York, NY, 1998.

[5] E. Chang, K. Goh, G. Sychay, and G. Wu. Cbsa: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):26–38, 2003.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, (16):321–357, 2002.

[7] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *Proceedings of ACM Multimedia Conference*, pages 540–547, 2004.

[8] K. Goh, E. Chang, and B. Li. Using one-class and two-class svms for multiclass image annotation. *IEEE Transactions on Knowledge and Data Engineering*, 17(10):1333–1346, 2005.

[9] K. Goh, B. Li, and E. Chang. Semantics and feature discovery via confidence-based ensemble. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 1(2):168–189, 2005.

[10] M. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, (6):429–449, 2002.

[11] X. Y. Liu, J. Wu, and Z. H. Zhou. Exploratory under-sampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 2009.

[12] M. R. Naphade, L. Kennedy, J. R. Kender, S. F. Chang, J. R. Smith, P. Over, and Hauptmann A. A light scale concept ontology for multimedia understanding for trecvid 2005. Technical report, IBM, 2005.

[13] K. Sande, T. Gevers, and C. Snoek. A comparison of color features for visual concept classfication. In *ACM Iternational Conference on Image and Video Retrieval*, 2008.

[14] A. F. Seaton, P. Over, and W. Kraajj. Evaluation compaigns and trecvid. In *Proceedings of the ACM SIGNMM International Workshop on Multimedia Information Retrieval*, pages 321–330, Santa Barbara,USA, 2006.

[15] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders. The challenge problem of automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*, pages 421–430, Santa Barbara,USA, 2006.

[16] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, Alaska, 2008.

[17] L. Wei, S. Maosong, and H. Christopher. Multi-modal multi-label semantic indexing of images based on hybrid ensemble learning. *Advances in Multimedia Information Processing PCM 2007, Lecture Notes in Computer Science, Springer Berlin / Heidelberg*, pages 744–754, 2007.

[18] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.