Comparison of Balancing Techniques for Multimedia IR over Imbalanced Datasets

Pablo Bermejo*, Frank Hopfgartner†, José A. Gámez*, José M. Puerta Callejón* and Joemon M. Jose†
*University of Castilla–La Mancha
Albacete, Spain
Email: {pbermejo, jgamez, jpuerta}@dsi.uclm.es
†University of Glasgow
Glasgow, United Kingdom
Email: {hopfgarf, jj}@dcs.gla.ac.uk

***Abstract*—A promising method to improve the performance of information retrieval systems is to approach retrieval tasks as a supervised classification problem. Previous user interactions, e.g. gathered from a thorough log file analysis, can be used to train classifiers which aim to inference relevance of retrieved documents based on user interactions. A problem in this approach is, however, the large imbalance ratio between relevant and non-relevant documents in the collection. In standard test collection as used in academic evaluation frameworks such as TREC, non-relevant documents outnumber relevant documents by far. In this work, we address this imbalance problem in the multimedia domain. We focus on the logs of two multimedia user studies which are highly imbalanced. We compare a naïve solution of randomly deleting documents belonging to the majority class with various balancing algorithms coming from different fields: data classification and text classification. Our experiments indicate that all algorithms improve the classification performance of just deleting at random from the dominant class.**

I. Introduction

In recent years, the rapid development of tools and systems to create and store private video enabled people to build their very own video collections. Besides, easy to use Web applications such as YouTube and Dailymotion, accompanied by the hype produced around social services, motivated many to share video, leading to a rather uncoordinated publishing of video data [1]. Despite the ease with which data can be created and published, the tools that exist to organise and retrieve are insufficient in all terms (effectiveness, efficiency and usefulness). Hence, there is a growing need to develop new retrieval methods that support the users in searching and finding videos they are interested in. However, video retrieval is affected by the Semantic Gap [2] problem, which is the lack of association between the data representation based on the low-level features and the high-level concepts users associate with video.

Approaches to bridge this gap are to select [3] or construct [4] features (audiovisual or textual) that can be used to form an optimal search query. In [5], we consider this relevance prediction problem as a supervised classification task. We aim at predicting the relevance of a multimedia document to a given search task by using different features to represent this document. The data collection which was used within this study consisted of many non-relevant and only a few relevant entities though. In order to train a classifier on an equal number of relevant and non-relevant documents, it is therefore required to exclude non-relevant documents from the training set. In this paper, we aim at comparing various techniques designed for balancing training sets in order to improve the algorithms classification success.

The paper is structured as follows. In Section II, we introduce work related to multimedia retrieval and balancing techniques. Then, we introduce our methodology in Section III and present our experiments in Section IV. Finally, we conclude in Section V.

II. Related Work

Our work is embedded into the context of multimedia retrieval and preprocessing of training sets for automatic classification. In this section, we briefly introduce problems in the field of multimedia retrieval and further introduce state-of-the-art approaches to solve the problem of imbalanced data sets.

Multimedia Information Retrieval is a broad description of various retrieval modalities, such as image retrieval, video retrieval, music retrieval and cross-media retrieval. Different from text retrieval, documents in a multimedia collection are full of feature representations. These features represent different aspects, e.g. visual features of an image, and are therefore predetermined to be used as content-based retrieval source. Even though large improvements have been achieved to index these features on a large scale, content-based retrieval is facing the Semantic Gap [2]. Multimedia objects sharing similar low-level features do not necessarily represent similar concepts.

One approach to bridge this gap is to exploit user feedback to identify documents that match the user's interests. Different types of feedback have been studied, the most prominent being explicit (e.g. [8]) and implicit (e.g. [9]) relevance feedback. When giving explicit feedback, users are asked to judge the relevance of a given document. However, users tend not to provide sufficient feedback [10], which is a big drawback in the approach. Another well-studied approach is to gather relevance evidence by analysing users' interactions with the

retrieval interface. This approach of gathering implicit relevance feedback is based on the assumption that users will show different interaction patterns while interacting with relevant and non-relevant documents, respectively.

A big challenge is to identify the optimal document representation or to select the best features for a given query [11] which can then be used to exploit this feedback (e.g. [12]). A novel approach of exploiting implicitly gathered data is to train a classifier of this implicit relevance feedback with the relevance of each document being used as class parameter. A problem, however, is that standard IR test collections consist of a large number of non-relevant and only a rather small number of relevant documents [13]. Hence, user log files will contain an imbalanced number of feedback given on according documents, a well-known problem in data classification.

Chawla et al. [14] differentiate between *data-level* and *algorithm-level* approaches to solve the problem of imbalanced data collections. *Data-level* methods perform modifications on the dataset before building a classifier [6]. These modifications consist of re-balancing the dataset by applying some techniques which have the advantage of controlling the data used for classification. However, this approach may lead to expensive computations. *Algorithm-level* approaches are classifier-specific and aim at biasing the training set, e.g. by giving weights to different classes in the training process [15] or adjusting the misclassification costs to more realistic values. Other approaches are to train a model using only the samples tagged with the minority class [16] or to apply classifier-specific techniques.

Chawla et al. [6] introduce the well-known *SMOTE* approach where a combination of over- and under-sampling is presented, resulting in an accuracy improvement for the minority class.

In [5], we aimed at improving the retrieval performance by comparing different kind of features to represent shots in a video collection. We used these features for a supervised classification and performed a balancing technique prior training the classifiers. This method (to which we refer to as *Alpha* method) randomly removes as many non-relevant documents as necessary in order to have the same cardinality for relevant and non-relevant documents. In this work, we use the same kind of features and we focus on comparing several balancing algorithms: *Alpha*, *SMOTE* and *distribution-based* methods. Besides we add a baseline composed of results from classification without balancing.

III. Methodology

A. Design of Experiments

In [5], we exploited log files of two user studies [17] to construct two datasets which were used to train a classifier for relevance judgement: the first constructed dataset consists of Behaviour Features as suggested by Agichtein et al. [4], containing continuous values. The second dataset consists of Vocabulary Features, containing a bag-of-words where each work is an integer representing a frequency. We created four datasets for each feature type, hereby denoted *Topic 1*, *Topic 2*, *Topic 3* and *Topic 4*, respectively. All datasets contain a binomial class $\{$relevant, non-relevant$\}$, which is highly imbalanced.

In this work, we aim at testing six different balancing methods as a preprocessing step before classification. Thus, we compare the performance of three different approaches: *Alpha*, as introduced in [5], performs under-sampling by randomly removing non-relevant documents from the collection; the well-known *SMOTE* algorithm, which is a state-of-the-art balancing method in data classification; and the distribution-based methods, which perform both oversampling and under-sampling and, depending of the used probability distribution gives rise to different algorithms [7].

We evaluate both the imbalanced and balanced datasets by performing a 5×2 CV with three different classifiers: NBayes, SVM and kNN. In the case of Vocabulary Features (see Section IV-A), we replace NBayes with Multinomial NBayes since this is the recommended classifier in literature [18] for text documents. Our main interest is two-fold:

1) analyse if balancing helps to outperform classification using both an imbalanced dataset and a random-based balanced dataset
2) statistically assess which of the compared balancing methods is more suitable for the multimedia IR problem

Since we are dealing with skewed datasets, Accuracy is not an appropriate metric to measure the performance of the classifiers [19]. So the selected metrics are Precision and F_1-measure of documents tagged with the minority class (relevant documents).

Finally, we compute statistical significance by performing a Wilcoxon signed rank test [20] with Conf. Level $= 95\%$ to compare all methods. Two sets of twelve values are used for the test. Each set is the joined output of the three used classifiers over each of the four topics.

B. Balancing methods

As mentioned above, we compared three different kind of balancing methods: *SMOTE*, *Alpha* and *distribution-based* methods.

SMOTE is a state-of-the-art algorithm which over-samples new instances of the minority class by creating new samples in the vector space between a minority class instance and one of its k nearest neighbours. Re-sampling of minority class documents is performed until the increasing minority class cardinality reaches the overall percentage as defined by the user. Formally, the sampling of a synthetic instance using the *SMOTE* method can be stated as follows: Let C be the binomial class such that $C = \{$relevant, non-relevant$\}$ and a set of instances $D = \{d_1, d_2, ...d_{|D|}\}$ such that each instance d_i is represented by N features $d = \{f_1, f_2, ...f_n, c_N\}$. For any document d_i belonging to the minority class c_m (relevant in this case), randomly chose another document d_j such that it is one of the k nearest neighbours of d_i and compute $subs = d_i - d_j$. Then, a new document d_k is generated by computing $d_k = subs \times random() + d_i$, where along all these operations documents are treated as vectors. Class parameters

are not taken into account. Chawla et al. [6] suggest a random under-sampling process of the majority class, which we also applied in this work.

The *Alpha* method [5] refers to a parameter α that indicates the percentage of the difference between cardinality of classes to be reduced. Formally: Let r = cardinality of relevant documents and nr = cardinality of non-relevant documents. Computing $diff = nr - r$, the number of non-relevant documents to be randomly removed is $rem = (\alpha/100) \times diff$. Thus, $\alpha = 100$ will result in a balance ratio (1:1).

Distribution-based methods perform several tasks on the training set. Given an input parameter P:

1) Re-sample minority class documents until reaching the predefined cardinality P.
2) Randomly under-sample majority class documents.
3) Balance Training set to a ratio $(1:1)$.

Formally, we can state the distributions-based methods as follows: For the minority class c_m (relevant in this case), a distribution is learnt for each pair f_i-c_m from all instances labeled with class c_m in the training set. Thus, new instances for class c_m are sampled by generating values for each feature f_i from the learnt distribution for f_i-c_k. Whenever the generated value for f_i is less than 0, it is set to the minimum 0 since negative values would not make sense for our features.

In this work, we instantiate the distribution-based method to the four probability distribution previously examined in text-classification: Uniform [7], Gaussian [7], Poisson and Multinomial.

IV. Experiments

In order to evaluate the previously introduced balancing methods, we exploit the log files of two different user studies [17]. In both studies, users were asked to interact with multimedia information retrieval systems and retrieve as many results as possible for four pre-defined search tasks. In both evaluations, the TRECVid 2006 video collection [21] was used where news video shots, the atomic unit of retrieval, are indexed based on the output of an automatic speech recognition system. A shot is defined as a part of the broadcast that has been created by a continuous recording from a single camera. The log files contained every key stroke and mouse click which was performed during this evaluation. They hence contain information how the participants of both studies interacted with relevant and non-relevant shots while using the system. In this section we first introduce the log files and then discuss the results after running the experiments introduced in Section III-A.

A. Datasets

For every search task, we converted the log file format to two different dataset representations. In the first dataset, we focus on behaviour features (see Table I) such as a click to start playing the video shot or the playing duration. Therefore, we represent every behaviour feature together with the shot that the action was performed on in the dataset. The actual relevance of the shot to the given search topic is defined as the class. Table II provides an overview of this dataset.

TABLE II
Description of Dataset using Behaviour Features.

	#Features	#Instances	Imbalance
Topic 1	13	5011	1:19
Topic 2	13	4542	1:16
Topic 3	13	4545	1:13
Topic 4	13	4701	1:40

In the second dataset, we focus on vocabulary features. Therefore, we represent every shot in the log files together with the terms that are aligned to it. Some shots in the corpus do not contain any speech. Thus, no transcript is available for these shots. We therefore ignore these shots, which results in fewer instances in the dataset in comparison to the behaviour feature dataset. Again, the shot's relevance was used as class. Table III provides an overview of the vocabulary feature dataset.

TABLE III
Description of Datasets using Vocabulary Features.

	#Features	#Instances	Imbalance
Topic 1	12957	2544	1:38
Topic 2	12957	2496	1:49
Topic 3	12957	2088	1:27
Topic 4	12957	2400	1:16

The last column of these tables depicts the imbalance ratio for each dataset. A ratio of (1:*n*) means that for each relevant document, the dataset contains n non-relevant documents. We can see that Topic 4 is the most skewed dataset for both shot representations. Therefore, we expect that the imbalance problem affects classifiers more aggressively in this case.

B. Settings

The balancing algorithms under study in this paper rely on various parameter settings which will be introduced in the remainder of this section.

SMOTE uses the default parameters in Weka except for the percentage of minority class cardinality, which is set as the necessary percentage to get P minority class documents. Then, random under-sampling of the majority class is performed until P majority class documents remain. Distribution-based methods *Uniform, Gaussian, Poisson* and *Multinomial* need the input parameter P. If, for example, $P = 500$ then, for each training set, 500 relevant documents will be sampled from the corresponding distribution and non-relevant documents will be uniformly deleted until the training set contains 500 non-relevant documents.

The *Alpha* approach needs parameter α which indicates the percentage of non-relevant documents to be randomly removed from the training set. We have set $\alpha = 100$, what means that we remove as many non-relevant documents as necessary to have the same number of relevant and non-relevant documents.

Since our aim is not to oversample non-relevant documents but to create training sets with the same number of relevant and non-relevant documents, we cannot set P to a higher value

TABLE I
BEHAVIOUR FEATURES USED TO REPRESENT VIDEO SHOTS.

Feature name	Description
ClickFreq	Number of mouse clicks on shot
ClickProb	*ClickFreq* divided by total number of clicks
ClickDev	Deviation of *ClickProb*
TimeOnShot	Time the user has been performing any action on shot
CumulativeTimeOnShots	*TimeOnShot* added to time on previous shots
TimeOnAllShots	Sum of time on all shots
CumulativeTimeOnTopic	Time spent under current topic
MeanTimePerShotForThisQuery	Mean of all values for *TimeOnShot*
DevAvgTimePerShotForThisQuery	Deviation of *MeanTimePerShotForThisQuery*
DevAvgCumulativeTimeOnShots	Deviation of *CumulativeTimeOnShots*
DevAvgCumulativeTimeOnTopic	Deviation of *CumulativeTimeOnTopic*
QueryLength	Number of words in current text query
WordsSharedWithLastQuery	Number of equal words in current query and last query

of the cardinality of non-relevant documents in any training set in the cross validation process. For example, *Topic 1* using *Behaviour Features* is a dataset with 5011 instances where about 4750 are non-relevant. Thus, since we perform stratified 5×2 CV, each fold will have $4750/2$ non-relevant documents, requiring P not to be greater than that value. Since we wanted to use the same P value along the four topics compared, the maximum value used in our experiments is 2000 for Behaviour Features datasets and 1000 for Vocabulary Features datasets.

C. Results and Discussion

Table IV shows the mean precision P and F_1-measure for classifiers NBayes, SVM and kNN after training the imbalanced datasets using two different kind of predictive features for video shots: Behaviour and Vocabulary features. We use these results as the baseline run to compare our results with.

TABLE IV
PRECISION (P) AND F_1-MEASURE FOR RELEVANT DOCUMENTS IN IMBALANCED DATASETS.

	Behaviour		**Vocabulary**	
	P	F_1	**P**	F_1
Topic 1	.104	.142	.050	.045
Topic 2	.114	.155	.041	.029
Topic 3	.137	.166	.052	.054
Topic 4	.089	.124	.000	.000
Mean	**.111**	**.147**	**.036**	**.032**

Tables V and VI list the computed precision and F_1-measure using Behaviour and Vocabulary features representation for video shots and running balancing techniques over the datasets prior to classification. In the remainder of this section, we discuss the outcome of results using, for each kind of features, the maximum P value.

Balancing Vs. Non Balanced. In terms of precision, we conclude that *SMOTE* and distribution-based (*Uniform*, *Gaussian*, *Poisson* and *Multinomial*) balancing methods statistically outperform the baseline using the Behaviour Features datasets. In the case of Vocabulary Features datasets both Uniform and Gaussian distributions are an exemption though. Balancing method *Alpha* does not outperform the baseline, what could indicate that randomly under-sampling the training set alone is not effective. In terms of F_1-measure, distribution-based methods under-perform. We interpret this in such as distribution-based methods increase Precision while losing Recall.

Distribution-based Vs. SMOTE. Using the behaviour features dataset, the *Gaussian* method significantly outperforms *SMOTE*, while the *Multinomial* method outperforms *SMOTE* using the vocabulary features datasets. This can be explained by the fact that the *Multinomial* distribution is designed to sample new text documents from existing ones. Besides, Behaviour Features are continuous values which cannot be modelled with this distribution, while the Gaussian Distribution is a quite general model which fits well to this dataset. *SMOTE*-balanced classifiers, however, result in a better F_1-measure.

Value for P. We have performed balancing of training sets by transforming them into datasets with the same cardinality (P) for both positive and negative classes (see Table VII). We have found that the larger P is, the better *SMOTE* and distribution-based methods perform in terms of Precision. Besides, non-random distribution-based methods improve most with increasing P value. As mentioned above, P is limited by the majority class cardinality so the maximum possible value for P was 2000 for Behaviour Features datasets and 1000 for Vocabulary Features datasets. For future works where ratio (1:1) is not fixed, a larger study of P influence would be of interest.

V. CONCLUSIONS AND FUTURE WORK

When users use an information retrieval system, their implicit actions while interacting with the system can be exploited to predict relevance of documents. As has been shown [5], this feedback can be used to train a supervised classifier that effectively predicts such relevance. Considering classical information retrieval experiences, users will by far interact more with non-relevant rather than relevant documents. Thus, the main problem of using such feedback data is that they are highly imbalanced. In this paper, we addressed this problem by evaluating various balancing methods.

We have evaluated the performance of six balancing methods: state-of-the-art *SMOTE* (directed over-sampling and random under-sampling), *Alpha* (random under-sampling) and

TABLE V
PERFORMANCE OF BALANCING METHODS IN BEHAVIOUR FEATURES DATASETS ($P = 2000$).

	Alpha		**Uniform**		**Gaussian**		**Poisson**		**Multinomial**		**SMOTE**	
	P	F_1	**P**	F_1	**P**	F_1	**P**	F_1	**P**	F_1	**P**	F_1
Topic 1	.081	.136	.125	.076	.215	.111	.081	.046	.088	.043	.113	.168
Topic 2	.092	.159	.164	.077	.212	.162	.131	.094	.141	.098	.114	.173
Topic 3	.126	.207	.268	.078	.191	.172	.239	.113	.282	.107	.155	.229
Topic 4	.065	.116	.154	.122	.155	.142	.093	.062	.083	.063	.100	.161
Mean	**.091**	**.154**	**.178**	**.088**	**.194**	**.147**	**.136**	**.079**	**.149**	**.078**	**.121**	**.183**

TABLE VI
PERFORMANCE OF BALANCING METHODS IN VOCABULARY FEATURES DATASETS ($P = 1000$).

	Alpha		**Uniform**		**Gaussian**		**Poisson**		**Multinomial**		**SMOTE**	
	P	F_1	**P**	F_1	**P**	F_1	**P**	F_1	**P**	F_1	**P**	F_1
Topic 1	.028	.053	.015	.019	.076	.037	.051	.060	.053	.063	.044	.065
Topic 2	.024	.045	.018	.021	.016	.022	.055	.057	.082	.074	.044	.066
Topic 3	.043	.080	.020	.020	.043	.053	.084	.093	.098	.084	.081	.112
Topic 4	.006	.012	.000	.000	.000	.000	.005	.008	.004	.006	.001	.002
Mean	**.025**	**.047**	**.013**	**.015**	**.034**	**.028**	**.049**	**.054**	**.054**	**.059**	**.042**	**.061**

TABLE VII
PRECISION IN DISTRIBUTION-BASED AND SMOTE METHODS AS P INCREASES

	Gaussian		**Poisson**		**Multinomial**		**SMOTE**	
P	Behav	Vocab	Behav	Vocab	Behav	Vocab	Behav	Vocab
500	0.145	0.025	0.104	0.041	0.107	0.046	0.102	0.039
1000	0.170	0.034	0.118	0.049	0.131	0.059	0.109	0.042
2000	0.194		0.136		0.149		0.121	

four distribution-based methods: *Uniform*, *Gaussian*, *Poisson* and *Multinomial* (directed over-sampling with replacement and random under-sampling). Our analysis (of logs obtained from interactive video retrieval user studies) suggest that balancing training sets using distribution-based methods result in a higher Precision in comparison to the other methods. More precisely, the *Gaussian* Distribution method provides the best balancing for continuous features (Behaviour Features) while the *Multinomial* Distribution method is best for text-based features (Vocabulary Features).

As future work, we propose to search for optimum ratios of (relevant : non-relevant) documents instead of fully balance training sets to ratios (1:1).

ACKNOWLEDGEMENTS

This work has been partially supported by the JCCM under project PCI08-0048-8577, MEC under project TIN2007-67418-C03-01 and FEDER funds. The second and fifth authors were supported by the European Commission under contract SALERO (FP6-027122). It is the view of the authors but not necessarily the view of the community.

REFERENCES

[1] S. J. Cunningham and D. M. Nichols, "How people find videos," in *Proceedings of JCDL '08*. New York, NY, USA: ACM, 2008, pp. 201–210.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[3] X. Geng, T.-Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," in *Proceedings of SIGIR '07*. New York, NY, USA: ACM, 2007, pp. 407–414.

[4] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," in *Proceedings of SIGIR '06*. New York, NY, USA: ACM Press, 2006, pp. 19–26. [Online]. Available: http://portal.acm.org/citation.cfm?id=1148170.1148177

[5] P. Bermejo, H. Joho, J. M. Jose, and R. Villa, "Comparison of feature construction methods for video relevance prediction," in *Proceedings of MMM '09*, 2009, pp. 185–196.

[6] N. V. Chawla, K. W. Bowyer, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[7] P. Bermejo, J. A. Gámez, and J. M. Puerta, "Improving knn-based e-mail classification into folders generating class-balanced datasets," in *In Proceeding of IPMU'08*, 2008, pp. 529–536.

[8] J. J. Rocchio, "Relevance feedback in information retrieval," in *The SMART retrieval system: experiments in automatic document processing*. Prentice-Hall, 1971, pp. 313–323.

[9] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: a bibliography," *SIGIR Forum*, vol. 37, no. 2, pp. 18–28, 2003.

[10] M. Hancock-Beaulieu and S. Walker, "An evaluation of automatic query expansion in an online library catalogue," *J. Doc.*, vol. 48, no. 4, pp. 406–421, 1992.

[11] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, vol. 30, no. 4, pp. 643–658, 04 1997.

[12] F. Hopfgartner, D. Vallet, M. Halvey, and J. M. Jose, "Search Trails using User Feedback to Improve Video Search," in *Proc. of the ACM Int. Conf. on Multimedia*, 10 2008, pp. 339–348.

[13] E. Vorhees and D. Harman, Eds., *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[14] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[15] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *In Proceedings of ICAI '00*, 2000, pp. 111–117.

[16] B. Raskutti and A. Kowalczyk, "Extreme re-balancing for svms: a case study," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 60–69, 2004.

[17] R. Villa, N. Gildea, and J. M. Jose, "A study of awareness in multimedia search," in *Joint conference on digital libraries*, 2008, pp. 221–230.

[18] A. McCallum and N. K., "A comparison of event models for naive bayes text classification," in *Learning for Text Categorization*, 1998, pp. 41–48.

[19] G. M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, 2004.
[20] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
[21] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06*, 2006, pp. 321–330.