Multimed Tools Appl
DOI 10.1007/s11042-008-0250-z

Using visual and text features for direct marketing on multimedia messaging services domain

Sebastiano Battiato · Giovanni Maria Farinella · Giovanni Giuffrida · Catarina Sismeiro · Giuseppe Tribulato

Abstract Traditionally, direct marketing companies have relied on pre-testing to select the best offers to send to their audience. Companies systematically dispatch the offers under consideration to a limited sample of potential buyers, rank them with respect to their performance and, based on this ranking, decide which offers to send to the wider population. Though this pre-testing process is simple and widely used, recently the industry has been under increased pressure to further optimize learning, in particular when facing severe time and learning space constraints. The main contribution of the present work is to demonstrate that direct marketing firms can exploit the information on visual content to optimize the learning phase. This paper proposes a two-phase learning strategy based on a cascade of regression methods that takes advantage of the visual and text features to improve and accelerate the learning process. Experiments in the domain of a commercial Multimedia Messaging

S. Battiato (✉) · G. M. Farinella · G. Giuffrida · G. Tribulato
Department of Mathematics and Computer Science,
University of Catania, Viale A. Doria 6, Catania 95125, Italy
e-mail: battiato@dmi.unict.it

G. M. Farinella
e-mail: gfarinella@dmi.unict.it

G. Giuffrida
e-mail: ggiuffrida@dmi.unict.it

G. Tribulato
e-mail: tribulato@dmi.unict.it

C. Sismeiro
Imperial College Business School,
Imperial College London, South Kensington, Campus,
London SW7 2AZ, UK
e-mail: sismeiro@imperial.ac.uk

Service (MMS) show the effectiveness of the proposed methods and a significant improvement over traditional learning techniques. The proposed approach can be used in any multimedia direct marketing domain in which offers comprise both a visual and text component.

Keywords Visual and text features · Learning in time and space constrained domains · Multimedia messaging services · Direct marketing

1 Introduction and background

The importance of today's direct marketing industry is reflected by its significant economic value [12]. In recent years, as a result of technological advances in computing and communications, new contact channels have become available. Beyond traditional channels, which include mail, catalog, and telephone contact, companies today can use multimedia channels as varied as email, mobile phone messaging, customized websites, addressable broadcasting, and direct-response TV and radio. It is this increase in the number of channels that has allowed, at least in part, the steady growth of direct marketing activities in recent years.

A common denominator across these direct marketing applications is the need to learn the *potential performance* of available offers. The objective is then to design better targeting and segmentation policies; such activities have always been at the core of the success of direct marketing campaigns [28, 30]. Traditionally, direct marketing companies have relied heavily on pre-testing to acquire knowledge and select the best performing offers [22].

1.1 Traditional pre-testing in direct marketing applications

The pre-testing process is simple and widely used across the industry. First, the set of offers under consideration is sent to a limited sample of potential customers. Then, depending on the sample response, companies compare the performance of each offer and select the best offer for each population segment. Finally, only the best offers are sent to the wider population of potential customers.

This is an approach that reduces waste, avoids sending irrelevant and potentially annoying messages to too many customers, and allows for higher performance and profitability. Indeed, traditional pre-testing has worked well for the many direct marketing applications characterized by a low cost of contact and a large customer base (e.g., traditional mail). Performance measures will vary depending on the specific application (catalogue, email, mobile messaging, etc.) and can include the number of items bought, the revenue per order, the click-through-rate[1] (CTR), and the number of calls generated.

[1] Click-through rate, or CTR, is a common way of measuring success for an advertising campaign targeted to mobile devices. For the scope of our paper it can be measured as the ratio between the number of users who clicked a specific offer over the total number of users that were exposed to that offer.

1.2 The new challenges of direct marketing applications

Recently, however, the broadening of the direct marketing process and the creation of new channels has brought new learning and knowledge acquisition challenges.

First, the number of new offers in need to be tested in most of these new applications is not only large, it also grows very fast. For example, in the context of mobile messaging it is not unusual to have more than 50,000 possible products or services to advertise at any moment, and the content catalogue can grow by twenty to thirty new items a day, a growth rate that is not likely to be reduced.[2]

Second, even though there are millions of customers to contact, the number of contact opportunities is small. For example, in our application mobile phone screens are not large enough to allow multiple offers to be easily visualized in a single message, and mobile phone operators impose limits on the number of commercial messages users can receive each day (receiving too many commercial messages a day increases the likelihood that a customer will cancel a service or switch operator due to annoyance). In addition, there are technical limits in current mobile targeting systems: though mobile phone companies can send the same message to millions of different customers, only few (thousands) of personalized messages can be sent on a single day. Hence, groups of customers will be exposed to the same commercial message, though different messages can be sent to different groups. This means that one single person can only be exposed to a very small fraction of all possible products and services.

Finally, many of the offers in new contact channels have very short life-spans, posing additional pressure to produce fast learning. Taking again the example of the commercial offers analyzed in the empirical section, these offers often expire close to the release date. In some cases offers need to be sent in a matter of days or even hours after they have been made available for selection and learning (e.g., news related services or holiday related offers), which further limits the opportunities and time for learning.

Hence, though it is essential for direct marketing applications to quickly understand customer's needs and interests, and to select the right services to promote at the right time and in the right way, current methods cannot deal with the challenges described above. The rapid growth in the number of offers to be tested, the limited number of learning occasions, and the reduced time available for learning, reduce considerably the effectiveness of the traditional pre-testing learning based approach.

[2] For mobile operators, sending commercial messages to their customers is very cost-effective: operators can easily reach millions of potential buyers at little cost, making the profit potential of these advertising-related services very high. In addition, in the case of mobile phone operators market saturation and fierce competition [23] have turned value added services (VAS), like the ones these commercial messages advertise, into significant revenue source and in some cases the only opportunity for revenue growth. Because these services are now central to profitability, mobile phone operators and independent production companies are becoming increasingly creative in generating and proposing new services and offers. The result is a rapidly growing set of possible services available.

1.3 Our contribution

In this paper we propose an approach to optimize the learning task under such learning constraints using a two-phase learning approach that applies a cascade of regression methods. The proposed approach takes advantage of the visual and textual features extracted from an offer multimedia content. We test the proposed approach in the context of a mobile marketing application in which a commercial multimedia message (MMS) is sent everyday to mobile phone users.[3]

The empirical results show that the proposed two-phase approach is significantly better than existing alternatives, and that the performance differential increases with the severity of the learning constraints. Hence, the main contribution of this research is to demonstrate that it is possible to make the learning phase more effective, with respect to traditional learning, by also analysing the visual and textual information present in each offer during the learning phase. The results point out also that each type of feature (visual and textual) contributes to the predictive power of the model, and that a Textons-based representation [15, 29, 36] achieves better results than the visual representation used in [4]. Finally, our results demonstrate that the proposed cascade of regression methods significantly improves system performance, as different regression methods are able to exploit more efficiently the information contained in each feature.

The remainder of the paper is organized as follows. Section 2 provides the details of the dataset and the MMS domain. Section 3 introduces the two-phase learning strategy, describes the process that have been employed to extract visual and text features from the MMS offers, and provides details on the regression tools used. Section 4 reports on the experiments and discusses the results in the MMS application domain. Finally, Section 5 concludes this paper with avenues for further research to improve learning under severe time and learning constraints.

2 The MMS direct marketing domain

The empirical application in this work (and in the test of the proposed approach) is based on a real direct marketing system that sends multimedia messages to mobile phone users. A multimedia message (MMS) is a special kind of mobile message with powerful capabilities. It may include text, graphics, and music and might allow some form of interactivity (today almost all new devices are designed to send and receive this type of messages).

In our application, a single commercial multimedia message is sent everyday to each user. Each message contains a commercial offer that advertises a specific product or service that can be purchased directly from the mobile phone with few clicks (e.g., a ringtone, a song, or a video).

In such a context, an optimization and targeting system that learns quickly and efficiently, selects the right message/product to be sent to each customer, and

[3] In the following section we explain in more detail what commercial mobile multimedia messages are and present several examples.

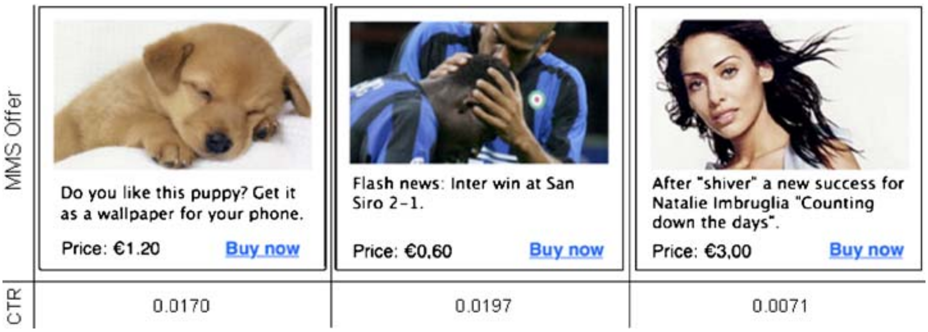

MMS Offer
Do you like this puppy? Get it as a wallpaper for your phone.
Price: €1.20
Buy now
Flash news: Inter win at San Siro 2-1.
Price: €0,60
Buy now
After "shiver" a new success for Natalie Imbruglia "Counting down the days".
Price: €3,00
Buy now
CTR
0.0170
0.0197
0.0071

Fig. 1 Some examples of available commercial MMS offers with resolution size 200×116. An image and short text is associated to each offer. Moreover, the considered dataset is labeled with a real click-through-rate obtained sending the offers with a traditional method during a period of 15 months

optimizes revenues, could provide a significant profitability boost. However, MMS targeting systems face the significant time and learning space constraints described above.[4]

One significant advantage of mobile operators for the learning task is that the current infrastructures keep detailed logs of all messages delivered and the response of each user. We can then track all messages and offers sent to customers (including their visual and text characteristics), and of the corresponding performance (e.g., whether the customer opened a message, viewed a page, bought a video, or clicked on a link). The information contained in these logs can then be used by an automated targeting system to aid message selection and customer targeting.

The proposed approach (which we present in detail in Section 3) has been tested on a real dataset of commercial multimedia messages collected from such logs. The messages were sent to mobile users in Europe over a period of 15 months. There were more than one million users who opted-in for the service and more than 70,000 possible direct marketing offers to advertise (only a subset of the 8600 items were labeled with a CTR due to the limit of the real targeting system in sending no more than twenty items per day).[5]

In this dataset, each commercial offer is composed by a small picture, a short description, and a price (Fig. 1). All images are encoded by JPEG standard with a high quality setting (i.e., no blocking is evident). The typical resolution size is 200 × 200 or 200 × 116. The overall dataset contains images with different types of pictorial content (e.g., face, people, buildings, cartoons, etc.) and the image contents are not always directly related to the category of the offer. Our database includes also the short text description associated with each offer. This text message contains on average twelve words and briefly describes the commercial offer though, unlike

[4] Given the speed of offer production in our application, even with daily contact (e.g., daily messages sent to mobile phone users), the number of offers to be tested grows at a faster pace than the rate at which a traditional pre-testing system is able to learn (while at the same time keeping enough potential customers for optimized delivery).

[5] The real targeting system could reach millions of users, but large segments of users would have to receive the same message. Only a maximum of 20 messages could be sent daily.

[2], the text in our application does not necessarily describe the offer content. It is the combination of image and text that provides the full information on the content. For instance, looking at the "Puppy MMS" image of Fig. 1 (first on the left), we are not sure what is exactly being sold unless we read the text associated with it. In other cases it could be the image more revealing on the content being sold than the text itself. That is also why we propose that the visual and textual content can each contribute to the prediction of the MMS performance independently. Finally, for each offer we also know the exact price charged to users who click and purchase the product/service and the real click-through-rate of each offer across all mobile phone users.

3 The two-phase learning approach

To improve learning on new offers under the time and learning space constraints described previously, we propose a two-phase methodology. Figure 2 presents the overall schema of our approach.

3.1 Approach overview

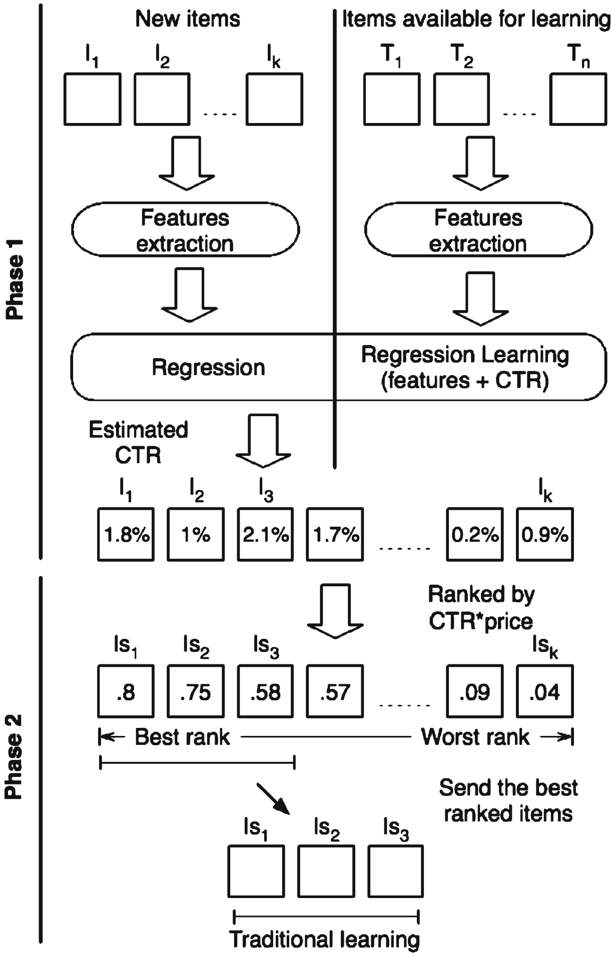
In Fig. 2 we depict our approach. The system first estimates how good each offer available for learning is likely to be (in terms of estimated CTR) by considering the static (non-behavioral) features and the past CTR of similar offers. The static features include the visual and text content of the offer. At the end of this first phase all new items are associated an estimated CTR. In a second phase, we constrain the learning, up to the system learning capacity, to the offers with the highest *performance potential*. The *performance potential* of each offer is defined as the product of the offer estimated CTR and its price, that is, we consider the expected revenue for each new delivered item (price is constant and known beforehand). Learning in this second phase takes place as in traditional direct market applications using the pre-testing method as discussed above: new offers are sent to the *learning panel* and the behavioral feedback (i.e., response) is registered and used for final offer selection.

Hence, in this two-phase approach, we apply the traditional learning only to a subset of the newly arrived items. Such subset is chosen based on the *performance potential* of each item, which is estimated taking into account its static features. Thus, we limit the traditional learning to the most promising items for which we have enough time to learn on. This in turn could potentially allow us to use a smaller sample of customers in the traditional learning phase and release customers to be included in the group subject to the more profitable optimized content delivery.

3.2 Exploiting visual and text features

The static features used to characterize the content of each MMS offer are extracted via automatic computer vision and text mining techniques. The extracted features and the known CTR are then used to train a regression-based method to estimate the CTR of new items (first phase) and to determine which offers should be subject

Fig. 2 The overall schema of the proposed two-phase learning approach



to further testing (second phase). In this section we describe the methods employed to extract the visual and text features.

3.2.1 Holistic encoding of visual content

Customers' evaluation of each offer is likely to depend on the interpretation of the offer visual content. We propose to rely on holistic visual features to model how appealing an offer might be for a user, and to use these visual cues in predicting a new offer CTR.

Studies in Scene Perception and Visual Search [6, 7, 27] emphasize that humans are able to recognize complex visual scenes at a single glance, despite the number of individual objects with different colors, shadows, and textures that may be contained in the scenes. Hence, a holistic representation to capture the visual content of the scene as whole entity is adequate in our application domain. Mobile phones have small screens and users tend to glance over images and not pay much attention to all details.

Recent studies from the Computer Vision research community [3, 16, 17, 24, 29] have efficiently exploited image holistic cues[6] to solve the problem of rapid and automatic scene classification, bypassing the recognition of the objects inside the scene. Our goal is to select a holistic representation able to capture the overall structure present in the MMS image. Because humans can process texture quickly and in parallel over the visual field [5], based on previous research, we considered texture as a good holistic cue candidate.

In this paper, to encode texture cues we first build a vocabulary of distinctive patterns, traditionally called visual words [18, 34, 37, 38], able to identify properties and structures of different textures present in an image [15, 29, 36]. Using the built vocabulary, an offer image is represented as an histogram of visual words. To build the visual vocabulary each image in the training set is processed with a bank of filters. All pixel responses are then clustered and the centroids for each cluster computed. These centroids, called Textons, represent the visual vocabulary. Each image pixel is then associated to the closest Textons taking into account its filter bank response. Hence, each image becomes a "bag" of Textons (or visual words) and we use the normalized histogram of Textons for each image as a descriptor for the holistic representation. The overall process is shown in Fig. 3.

We use the bank of filters suggested in [37] and the k-means clustering algorithm to build the Textons vocabulary. More specifically, each pixel in each image has been associated with a 17-dimensional feature vector obtained by applying three Gaussian kernels ($\sigma = 1, 2, 4$), four Laplacian of Gaussian kernels ($\sigma = 1, 2, 4, 8$) and two derivative of Gaussian Kernels ($\sigma = 2, 4$) on x and y directions. The Gaussian kernels have been applied on *Lab* channels whereas the remaining filters only on the *L* channel.

In Fig. 4 we present some images from the test dataset and the closest training images with respect to the distribution of Textons. The similarity between test and training images is computed using a metric based on the Bhattacharyya Coefficient [11] on the Textons-based representation discussed above. As it can be seen from the images in Fig. 4, there is a semantic consistence in visual content between the test images and the corresponding closest training images.

Because other types of visual information have been found to adequately capture appearance and to be powerful in predicting an offer performance (see [4] for more details), an alternative to this Textons-based representation was also considered. These alternative image descriptors include three types of visual information:

- *Colors*: full histograms and related statistics in the RGB, HSV, and CIE Lab color spaces;
- *Filter Response*: a complete set of band-pass filters to select both high frequency details and global appearance;
- *Semantic Analysis*: to take into account objects and/or scene context we use pixel detectors using a Bayesian classifier for different object classes [21] (e.g., Sky, Skin, Vegetation). The Bayesian classifier was learned using hand-labeled

[6] By definition, a holistic cue is one that is processed over the entire human visual field and does not require attention to analyze local features [29].

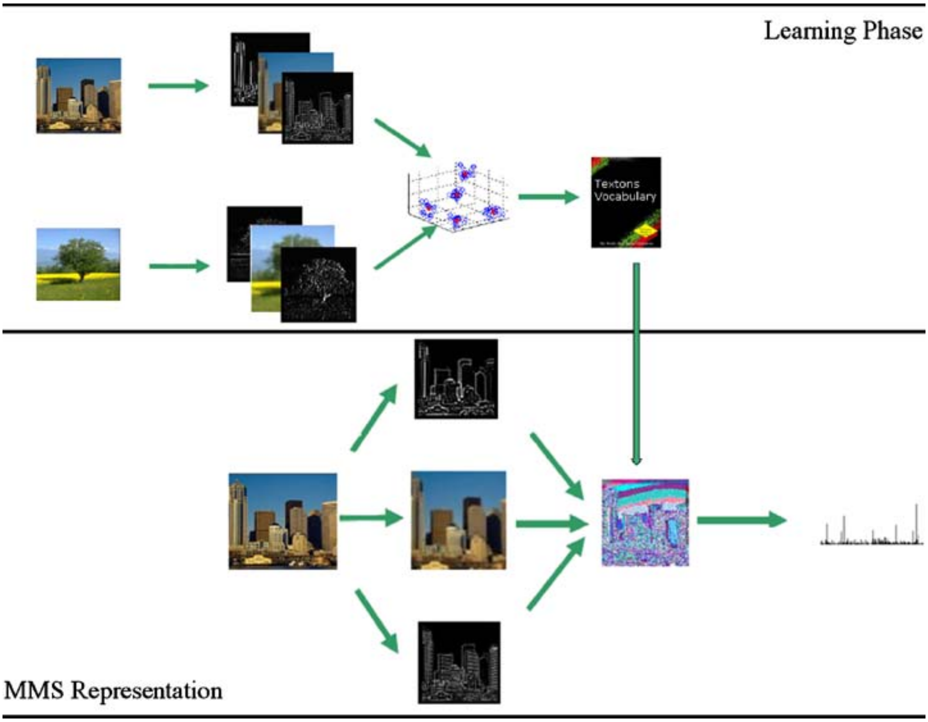


Fig. 3 *Learning Phase*: MMS images from the training set are convolved with a bank of filters. Filter responses are clustered into Textons using the K-Means algorithm to form the Textons vocabulary. *MMS Representation*: Given an MMS image, its corresponding representation is generated by first convolving it with the bank of filters and then labelling each filter response with the closest Texton within the vocabulary. The normalized histogram of Textons, i.e. the frequency with which each Texton occurs in the labelling normalized with respect to the total number of Textons in the labelling, forms the visual content descriptor of the MMS image

images (to define the color space of the classes) and the classification rule was a maximum a posteriori.[7]

Hence, the final feature vector of this alternative visual representation included color histograms, filter responses on color, and percentage of image pixels belonging to one of these the "appearance" classes from the Bayesian classifier.

We will compare the predictive power of this alternative visual representation and the Textons-based representation in the results section (Section 4).

[7] By using the Bayesian classifier one can infer the presence of faces in an image by the skin appearance in the pixel domain; likewise, an outdoor context can be inferred by sky and/or vegetation appearance [21]. We used these three types of visual information in our system as proposed by [4] and used the percentage of pixels belonging to each one of these appearance classes as determined by the Bayesian classifier to describe each image. The disadvantage of this method is that it required hand-labeling of a training set.

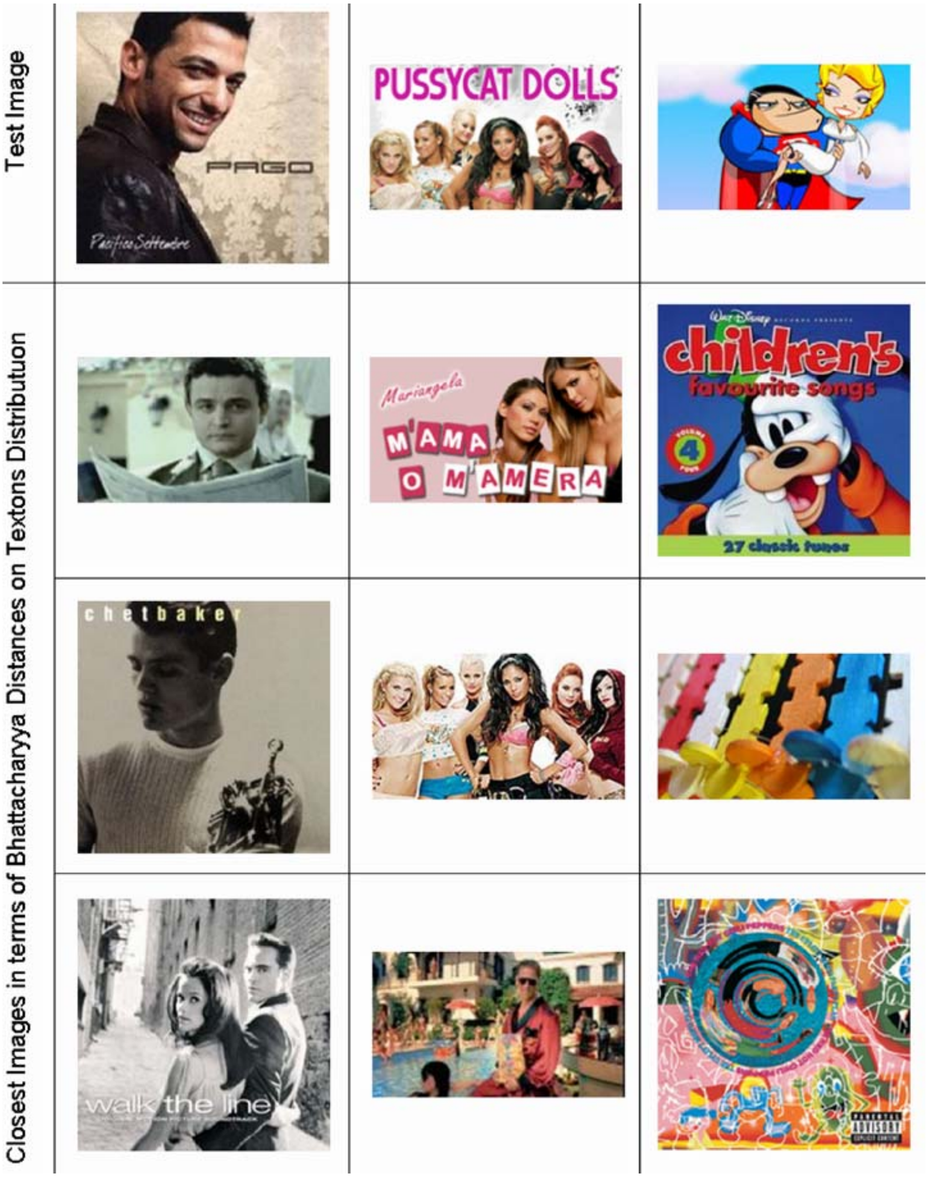


Fig. 4 Some test images are depicted at the *top row*. The closest training images in terms of Textons distribution are reported in the *second* and *third row*. The closest images are semantically consistent with the related test images

3.2.2 Extracting text features from offer description

To determine the static features of the short descriptive text message of each offer, we collected the set of most common words among all messages in the training set (i.e., among all messages for which performance is assumed known). All stop words and the least significant words of messages were removed from all words extracted. The significance of a word is function of its entropy, which is computed given the distribution of that word across all messages in the training set. Hence, words that appear on a large percentage of messages tend to be less discriminant (thus, less significant) than words appearing on fewer messages.

In addition, we removed also all words appearing in a very limited number of messages (less than five) because their likelihood of occurrence in the learning set is very small (thus, it is also unlikely that we will be able to use these words in the generalization phase). Stemming [14] was also considered but it did not improve the performance of the proposed approach.

After filtering the words from all offer descriptions, the learning space was dummified by creating a boolean variable for each word. Given a phrase, we set to one all boolean variables corresponding to the words composing the text, and to zero otherwise. This way we obtain a description of the text of each offer as a function of dummy variables

On average, five words per message have been extracted, and all of the words extracted have been used in the first learning phase. Moreover, the Term Frequency Inverse Document Frequency (TF-IDF) normalization schema has been considered on the text representation [25]. Specifically, TF-IDF was employed for Locally Weighted Regression method in which opportune metrics can be defined (see the following section for more details).

3.3 Learning methods

Our objective, in the first learning phase, is to predict the CTR for those items not yet tested in the population. Hence, the dependent variable of our first learning phase is the CTR, which has been measured only on items previously sent to mobile phone users (training set). For those messages not yet tested in the population, we use the *predicted* CTR to sort the offers and decide which ones to subject to further testing (i.e., to send to the second learning phase).

To perform the first learning phase, and predict the CTR, we use the observed CTR and the static features of the commercial offers previously sent to users (obtained through automatic computer vision and text mining techniques) to train a regression-based model. For each MMS three feature vectors were created:

- f_{Color}: vector representing color-based features for each offer (colors, filter response and image description from semantic analysis, as described above and in [4]);
- $f_{Textons}$: vector representing the probability distribution over the visual vocabulary (Textons) of each offer.
- f_{Text}: vector representing the text-based features for the short-text associated to each offer; the vector contains the dummy variables for the words present in the offer text.

Previous work has applied decision trees in a similar CTR prediction context [4]. However, decision trees require the dependent variable to assume a discrete form and, as result, the authors in [4] convert CTR into performance classes (e.g., Good, Fair, Bad). Unlike this previous work, we propose that it is possible to perform a more accurate and robust learning process by taking into account the continuous nature of the CTR variable. Hence, instead of decision trees, we propose to use regression-based methods that can better accommodate a continuous dependent variable.

Different regression models were tested on our dataset: Linear Regression (LR) [1], Regression Tree (RT) [8], Locally Weighted Regression (LWR) [10], and Support Vector Regression (SVR) [33]. These regression methods were selected after considering the properties of each type of static feature.

3.3.1 Linear regression

The simplest one is the LR approach, which is able to recognize the hyperplane that best fits the training data. Obviously, LR assumes a linear relationship between dependent and independent variables. Although LR performs poorly on many real dataset, we test this method as a benchmark.

3.3.2 Regression tree

The RT method has been previously used in similar prediction problems [4]. It derives a set of if-then logical (split) conditions and does not assume any linear relationship between dependent and independent variables. Considering the binary nature of text features, RT is an especially sensible choice for these features. However, one major drawback of RT models is that a threshold value must be determined in each stage of the tree construction and might not correspond necessarily to the optimal boundary of bipartition of the input subspace. In our experiments we implement the RT method using the WAGON library [35].

3.3.3 Locally weighted regression

Our CTR prediction problem relies on image and text-based features, making the problem highly complex and non-linear. The LWR is a memory-based method that can take advantage of the localized information at the neighborhood of each feature vector: it performs a regression around a point of interest using only data near that point. The basic assumption of LWR is that the neighborhood of each point in the dataset share similar values for the dependent variable. An important element in LWR estimation is the metric to be used to compare similarity of feature points. We note that, in using LWR, we have chosen a metric based on the Bhattacharyya Coefficient [11] for Textons, the L_2 metric for color-based representations, and the cosine similarity (dot product) and the TF-IDF methodology for text features. The LWR algorithm was implemented following the work of [10].

3.3.4 Support vector regression

Unlike the previous methods, the SVR method learns a nonlinear function in a kernel-induced feature space. Such a regression method, which takes into account the entire representation of an offer, could be more appropriate for visual representations based on Textons (we recall that Textons encode an image as a whole entity and provides a holistic representation of the information contained in the image). The two main disadvantages of this approach are the parameter tuning and the choice of adequate Kernel for the feature space. In our experiments we have used the epsilon-SVR algorithm [32] within the LIBSVM library [9], and a radial basis function kernel.

4 Experiments, results and discussion

In this section, we describe in detail the experimental setup of our simulated MMS direct marketing system. Then we describe the results of our experiments using the proposed learning approach and discuss our main findings.

4.1 Experimental setup

We randomly split the dataset described in Section 2 into three partially overlapping subsets of 4500 MMS offers each (see Fig. 5 for a depiction of the testing and training sets). By testing the proposed approach on three datasets we minimize the chance that our results depend on a "lucky" (or "unlucky") draw from the distribution of users and commercial offers (we note that as we only observe the mobile phone users and commercial offers associated to one single company in the market). In addition, because the three datasets cover different time periods we can also study whether the proposed approach performs differently over time.

For each dataset we applied the learning methods independently and simulated the MMS targeting system as we describe in the next section. We note also that though the three datasets are partially overlapping this does not pose a problem for the learning phase as the learning sets used in each simulation are all completely different (the learning sets per se do not overlap). Because the results across the three

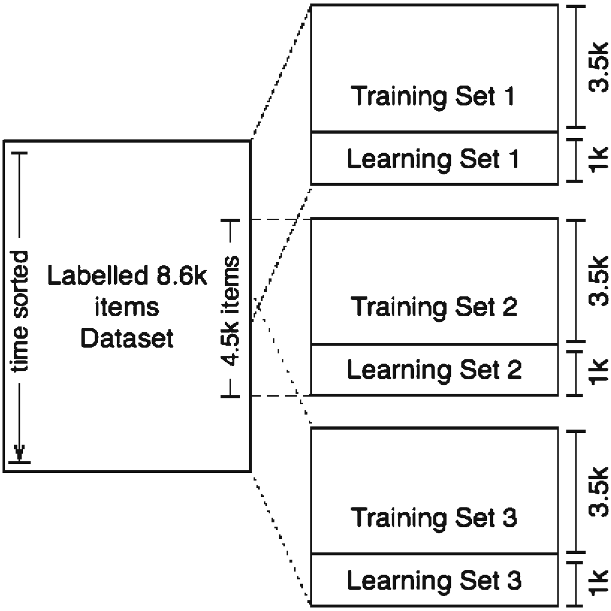


Fig. 5 The labeled dataset was subdivided into three subsets to perform the testing phase. Each subset was used to simulate the MMS direct marketing system. The results relative to each subset were collected and used to analyze the performances of the proposed two phase learning approach

datasets did not reveal any significant differences, we report the average performance across the three experiments.

The MMS offers were sorted based on their arrival date and time and the first (hence oldest) 80% of the offers within each subset were used as training data.

The remaining 20% of each dataset was retained as test data. The two-phase learning approach presented in Section 3 was used to select the best offers to be subject to further testing and then the performance of the MMS targeting system is measured based on the final selected offers. Performances of the simulated system are measured as described in Section 4.3.

For the Textons-based representation we tested for different visual vocabulary sizes: in the clustering of the visual features we tested for $k = 100, 200, 400$ where k is the number of clusters (i.e., the number of visual words). In the experiments presented in the following subsections, a vocabulary with 200 words was used because our preliminary tests revealed that this number provided the best results. We also tested for the number of dummy variables to use in representing the MMS text. The reported experimental results are based on the best number of dummy variables for each one of the methods applied. Specifically, through the filtering process discussed in Section 3.2.2, we obtain the best results using an average of about 1150 dummy variables in SVR and RT, and an average of 881 dummy variables in LWR. The overall parameters used in our experiments are reported in Table 1.

Table 1 Regression methods settings

Regression method	Parameter	Value
LR	Text features	Binary representation (1150)
	Visual features	Color based, textons based
RT	Text features	Binary representation (1150)
	Visual features	Color based, textons based
	Stop[a]	8
LWR	Text features	TF-IDF representation (881)
	Visual features	Color based, textons based
	Metric on text	Cosine similarity
	Metric on color	Euclidean L_2
	Metric on textons	Bhattacharyya coefficient
	Polinomial degree	0
	NN-bandwidth	39
	$K(d) = e^{-\frac{d}{\sigma}}$	$\sigma = 0.15$
SVR	Text features	Binary representation (1150)
	Visual features	Color based, textons based
	$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}$	$\gamma = \frac{1}{\text{\# features}}$

[a] Minimum number of examples for leaf nodes

4.2 The MMS targeting simulation system

Based on the real data from the MMS delivery system we set up a simulated delivery system to test the different approaches across multiple scenarios and conditions. We note that, given the number of people in the customer base, and the reduced number of opportunities to contact them, we are able to learn only on a small number of new MMS offers each day. We will simulate these conditions and allow for the severity of the learning constraints to vary.

In our simulation we define the *sending rate*, *S*, to be the number of daily (learning) trials. This corresponds to the maximum number of MMS offers we can send to the *learning panel* each day (the *learning panel* being a set of customers we use for learning on new offers). We define *arrival rate*, *M*, as the average number of new MMS items added to the offer catalogue each day. Note that, before a final decision is made on which MMS items to send to the optimization portion of the customer base, we should learn on the revenue potential of each one of these new items.

In the context considered, the *arrival rate* is greater than the *sending rate*. We call *overcapacity rate* the difference (*arrival rate–sending rate*). In addition, we call *overcapacity* the total number of unlearned messages. Hence, *overcapacity* measures the number of items on which we cannot learn, under the given learning space constraints. Since *overcapacity* grows monotonically each day (the rate of growth is given by the *overcapacity rate*), with a positive growth rate the learning task becomes more difficult over time.

Finally, we define *base size* as the number of offers that have never been sent to customers and are present at the beginning of the learning phase (in a sense, the *base size* is the *overcapacity* measured just before the first phase in our model takes place). As time goes by, if the *overcapacity rate* is greater than zero, the number of unlearned offers increases (i.e., the *overcapacity* grows). This leads also to a growth of the *base size* just before a subsequent learning phase is initiated.

The MMS targeting simulation system is assumed to run for 14 consecutive days. We set the *sending rate*, *S*, to 10 (the maximum number of daily trials) and varied the daily *arrival rate* M ($M = \{10, 15, 20, 25, ..., 50, 55\}$). For each value of the *arrival rate*, we assume it remains constant throughout the entire testing period. This means that, for each run, the *overcapacity rate* is constant and equal to $M - S$. Hence, each day, we accumulate $M - S$ new offers on which we are not able to learn. Finally, we set the initial offer catalogue size to zero (i.e., we start without any items requiring learning).

4.3 System assessment

Once the learning phase is complete, we can compute the overall expected performance of any regression method (*RM*), considering the set of the MMS offers chosen, as follows:

$$\text{Performances}_{\text{RM}} = \sum_{m \in \Omega} \text{CTR}_m \times \text{PRICE}_m, \tag{1}$$

where CTR_m represents the real click-through-rate of offer m, Ω represents the set of chosen offers when the regression method *RM* is used, and PRICE_m is the offer price (i.e., the cost for mobile users). (We note that in the specific domain we are studying the benefit for the mobile phone company accrued from selling each product or service is directly proportional to its price. The mobile phone company receives a fixed share of the price of what is sold and hence, modeling benefit or overall revenue will provide the same results.)

We can also compute the overall system performance when we randomly select the offers for testing as in the traditional learning case ($\text{Performance}_{\text{RAND}}$). Note that $\text{Performance}_{\text{RAND}}$ represents the performance lower bound from a traditional pre-testing learning system. To assess the performance of the proposed approach, a data mining measure called *Lift* is used:

$$\text{Lift} = \frac{\text{Performances}_{\text{RM}}}{\text{Performances}_{\text{RAND}}}. \tag{2}$$

The *Lift*, as defined above, measures how well the proposed approach performs in terms of expected revenue relative to the traditional pre-testing (i.e., without the offer pre-screen based on the estimated CTR using visual and text features). In addition to the *Lift* measure, and to better assess the performance of the proposed learning approach in estimating the CTR of each offer, we also compute the root mean squared error (*RMSE*) on the CTR predictions from the test data:

$$\text{RMSE}(\Psi) = \sqrt{\frac{\sum_{m \in \Psi} (\widehat{\text{CTR}}_m - \text{CTR}_m)^2}{|\Psi|}}, \tag{3}$$

where Ψ represents the test data.

RMSE is a frequently used measure of the distance between the predicted values of a dependent variable and the true variable values for a given prediction method. Note also that in Eq. 3, the CTRs values were normalized to lie in the range [0, 1000]. The *RMSE*(Ψ) and *Lift* measure is computed for each simulation run and then the average is taken across all datasets and all runs. Below, we will report on such averages.

4.4 Performance comparison when using different visual features

Previous research has demonstrated that color-based visual features, as described in Section 3.2, can significantly improve CTR prediction. Researchers in [4] report on several experiments showing that combining color-based and text features performs better than a system using text features alone. In this work we propose a holistic Textons-based representation that, we believe, can more adequately match the visual processing of mobile phone users. Hence, to better understand which visual representation, color- or Textons-based, best predicts the CTR of MMS offers, we compare how well the two representations predict CTR based on *RMSE*(Ψ).

It is also our goal to understand which regression method (RT, LWR, SVR) can better extract the information in the offer visual representation, and why differences of performance can be found. We report the *RMSE*(Ψ) for the different regression methods in Fig. 6.

As it can be seen from Fig. 6, the Textons-based visual representation outperforms color-based features when using LWR or SVR (the two best performing methods

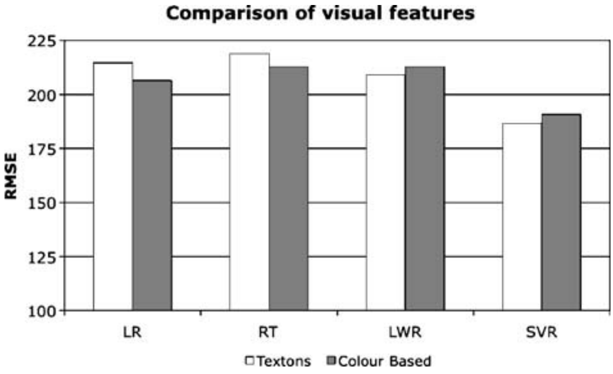


Fig. 6 Textons based vs color based

overall). As expected, the best results are obtained when using Textons and SVR. In fact SVR is able to extract the information contained on the global visual representation of Textons more efficiently. The color-based features perform better only when the RT method is used because this regression method discriminates by considering one single component feature at each level of the tree. So RT is not able to properly extract the information contained in the Textons-based visual representation.

We further note that SVR and LWR outperform RT independently of the image-related features used. This result allows us to conclude that Textons and SVR combined are better suited for our application. In the following results, our tests rely solely on Textons to capture an offer visual features.

4.5 The proposed two-phase approach relying only on Textons features

We have performed an additional test to understand whether our proposed two-phase approach outperforms traditional pre-testing when visual features are considered in isolation (i.e., without using text features). Figure 7 reports the *Lift* of our two-phase approach using SVR and Textons visual features.

As it can be seen in Fig. 7, our approach outperforms traditional pre-testing when visual features are used alone. The performance is better than the traditional approach already at low *arrival rate* values and increases in more constrained time and space domains. The improvement is more than 16% considering an arrival rate equal to 30 offers per day.[8] This means that, in direct multimedia marketing applications, we can effectively use the information on visual content to achieve better targeting results.

4.6 Performance comparison when using different types of features

We are further interested in testing how the different types of features (Textons- and text-based) contribute to CTR prediction and whether the combined use of these

[8] Taking into account the overall simulation settings, 30 offers per day is an arrival rate comparable to the mean arrival rate observed in the real system.

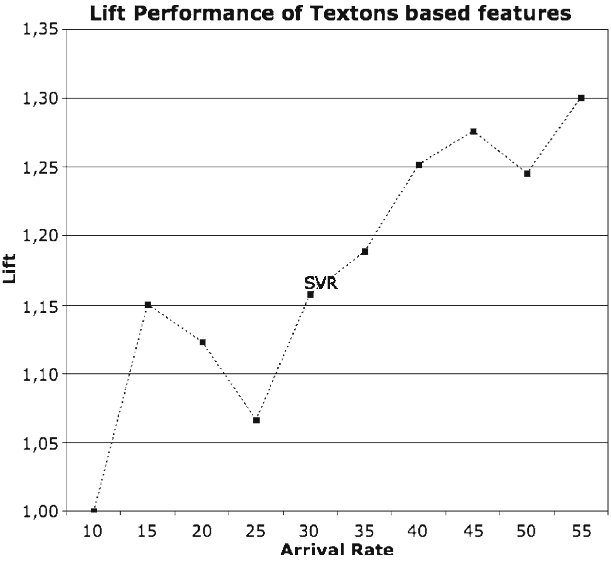


Fig. 7 The *Lift* results confirm that our approach outperform the traditional learning also using just Textons based features alone

features can improve targeting. In addition, we also want to determine the most appropriate regression method for each type of feature considered. Figure 8 provides a $RMSE(\Psi)$ comparison for each method-feature combination. This comparison allows us to determine which regression method best extracts knowledge from each feature (this turns to be a useful information when combining features in a regression cascade).

As it can be seen from Fig. 8, the SVR method provides the best results for Textons-based features (see our discussion of this result in Section 4.4) and the best method for text-based features is the LWR. Just like for Textons, note that RT is less powerful than the other two methods. Because RT is not able to consider word-related dummies as a joint entity, we can conclude that it is better to take into account different words together than to look at each word singularly to capture the semantic of an offer. Instead, LWR works well because the cosine similarity metric employed,

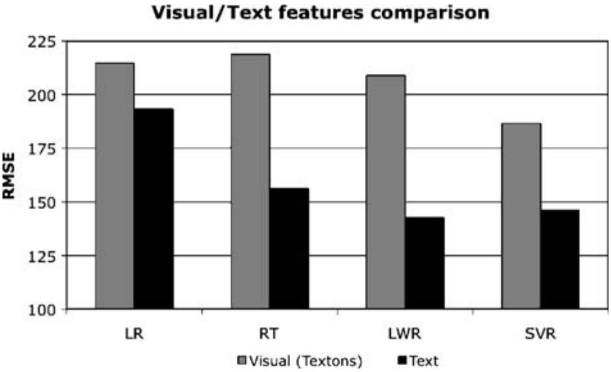


Fig. 8 $RMSE(\Psi)$ obtained with the three regression methods RT, LWR and SVR working on each kind of extracted features

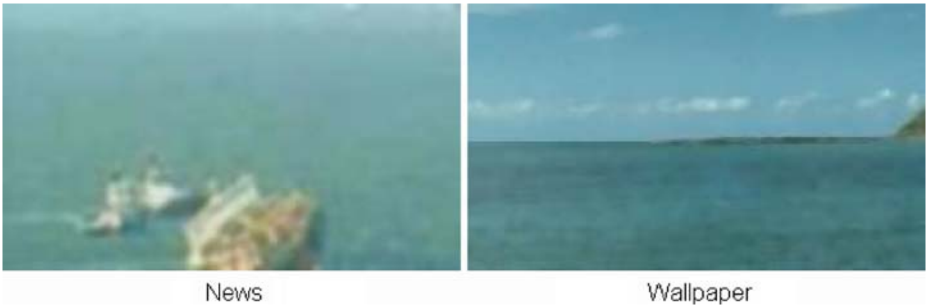


Fig. 9 Two images extracted from our dataset. The images are similar in terms of visual content, but they come from two different categories

combined with the TF-IDF representation, does exactly that: consider the joint role of all words.

Linear regression performs poorly in all cases and not will be further analyzed.

As shown in Fig. 8, we also find that text features perform better than visual features across all regression methods. One possible reason for this result is that the text might be working as a proxy for the offer category (i.e., telling us whether the offer is connected to music, sports, phone wallpaper, etc.). In this specific domain application, we do not know a priori the offer category (the classification of the offers into categories is not available to us). Hence, we associate the power of text variables in predicting CTR for each MMS to the ability of such short text in providing a category semantic. In contrast, the visual representation, though still showing predictive power, does not perform as well as the text. We conclude then that the visual component does not provide as clear cues as the text when predicting an offer category or its likeability. Indeed, many of the offers from different categories are very similar in terms of visual content (see Fig. 9 for two examples from very different categories).

We further tested this contention by combining the visual and text features using a SVR model and determining whether it provided a better fit than the same model using both types of features independently. Recall that SVR obtained good performances on the two types of static features (see Fig. 8). Table 2 reports the $RMSE(\Psi)$ results for the SVR model.

From Table 2 we can see that combining visual and text features provides better predictive accuracy than considering each feature independently, confirming that both sets of variables might be capturing different aspects of the offers likeability (the SVR model using the visual and text features together outperforms the SVR

Table 2 The $RMSE(\Psi)$ results of using visual and text features singularly and jointly in SVR are reported

Regression approach	RMSE(Ψ)
SVR(f_{Textons})	186.48
SVR(f_{Text})	146.35
SVR($f_{\text{Text}}, f_{\text{Textons}}$)	145.83

The combination of visual and text information gives the best results model when using each kind of feature separately). Hence, Textons-based features seem to be able to capture visual content whereas text-based features seem to add semantic information related to an offer category.

In sum, one important result from our experiments is that text mining on an offer short text might provide a good proxy for the offer category, and as result, might help predict its CTR. Our results further suggest that by combining the two sets of static features, visual and text-based, we are likely to obtain better predictions (as both sets of variables seem to capture different aspects of the message). In addition, these results also indicate that a greedy-like combination of regression methods could potentially provide additional prediction improvements.

4.7 Exploiting visual and text features together (cascade of regression methods)

Considering the result obtained with the regression methods described above, and taking into account the properties of the different MMS features (e.g., that Textons provide a global representation, and that text is represented by binary variables for each relevant word), we believe the combination of regression methods might be better suited to capture the properties of each static feature (visual and text) and could take into account the specific strengths of each regression method. Hence, what we propose is to use a combination of regression methods known as a regression cascade.

Note that SVR did not provide the best results for text features (Fig. 8). The best method to exploit text features was the LWR model. In addition, as discussed in the previous section, our findings seem to suggest that we should use text features before the visual component in order to capture the category of each MMS offer (e.g., music, sport, wallpaper, etc.). Then, we could add the visual features to discriminate between different visual content and improve predictive ability. This analysis induced us to use a greedy combination of features and regression methods. The final regression approach selected involved the LWR and the SVR models in cascade to take first advantage of the text features and then the visual features.

We tried two alternative formulations for the cascade of regressions. The first one combines the predicted CTR using text features and the LWR method with the visual features using a SVR method. The second alternative cascade tested was similar to the first cascade but it also included the text features directly in the SVR model (this

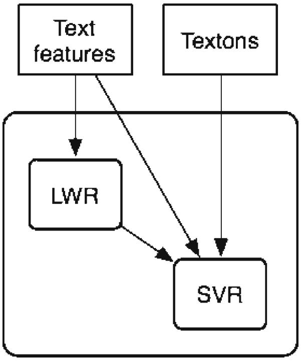


Fig. 10 The cascade of regression method involved in the first phase of our learning approach

Table 3 The $RMSE(\Psi)$ results of using visual and text features in SVR are reported in comparison with the results obtained by the proposed cascade approach

Regression approach	RMSE(Ψ)
SVR($f_{\mathrm{Text}}, f_{\mathrm{Textons}}$)	145.83
SVR(LWR(f_{Text}), f_{Textons})	144.04
SVR(f_{Text},LWR(f_{Text}), f_{Textons})	142.82

A cascade of regression methods achieved the best results

cascade model is presented in Fig. 10). We added text features again to the SVR model because both the SVR and the LWR models provided comparable result on these features, though they exploit the data in a very different manner (SVR looks at the global feature space whereas LWR looks locally around feature points). Table 3 reports the $RMSE(\Psi)$ results of these cascade of regressions.

From the analysis of Table 3 we conclude that the regression cascade improves the predictive ability of our system, estimating more accurately the CTR of each offer and improving offer targeting. The results seem also to suggest that a cascade of regressions can better capture the properties of each kind of representation (visual, text), and take into account the benefits of each regression method.

To better understand the performance of the overall system using the cascade approach, we computed the *Lift* for alternative model formulations. Figure 11 presents the *Lift* results. It is clear from the figure that using the proposed approach (irrespective of the final regression formulation used) improves significantly the overall performance of the system, producing significantly higher revenue. This improved performance is the result of the offer selection during the first learning phase. Because only the most promising offers are subject to further learning, we reduce the waste of scarce testing opportunities by discarding the weakest offers and not submitting these to further testing.

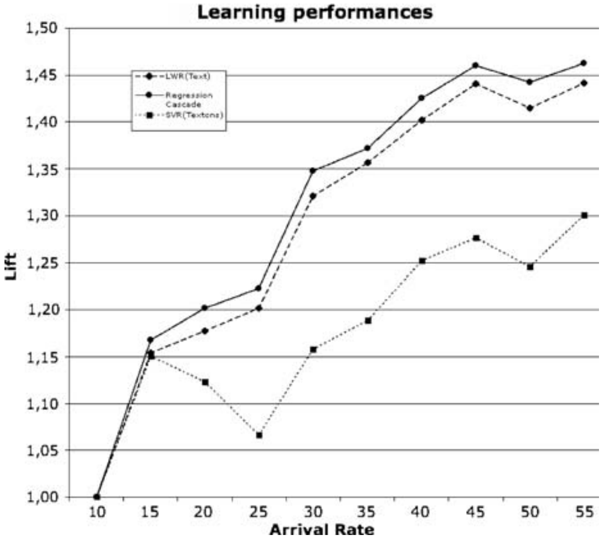


Fig. 11 The plot reports the *Lift* result obtained by using visual and text features singularly and by using the cascate of regressiom method

Figure 11 also shows that the direct marketing system outperforms the traditional learning approach more than 35% when the *arrival rate* is 30 items per day (we note that we are using only a subset of the offers of the real system in our training and testing procedure and an arrival of 30 items per day implies a similar mean *arrival rate* as in the real system). In addition, we can clearly see that system performance increases as the constraints become more severe (more than 30 items arrival per day) reaching 45% of improvement when we receive new 55 messages daily for testing. It is interesting to note also that the proposed approach outperforms the traditional learning approach also when using only visual features. In such case, the proposed approach outperforms the traditional testing in more than 16% when the *arrival rate* is 30 offers per day.

5 Conclusions and future work

A successful approach to improve performance of direct marketing systems on new multimedia channels will require the contribution of a wide range of disciplines and technologies including computer vision, data mining, statistics, and marketing. In this paper, we propose a two-phase learning approach for a direct marketing application subject to severe time and space learning constraints. We test our approach using data from a Multimedia Messaging Service (MMS) that delivers product and service offers to mobile phone users daily. Our approach exploits the visual and text features of each MMS offer through a cascade of regression algorithms that estimates the potential (in terms of expected revenue) of each offer. We then test the performance of only the best offers using traditional pre-testing methods.

Our results demonstrate that the proposed approach leads to a considerable improvement in overall performance even when it relies only on visual features. This means that when visual information (e.g. images) are present, it is possible to take advantage of visual content to make the learning phase more efficient and effective. Hence, researchers and businesses could use the proposed approach in other domains in which visual and text information is available (e.g. offers sent by email, interactive television, etc).

Future work could evaluate online techniques for the learning phase strategy (e.g. online boosting [26, 31]), and could take into account other active learning strategies based on uncertain sampling. In addition, customers' behavior and the offer price could be exploited as features and be used jointly with visual and text features to predict performance. Moreover, new visual representations taking into account priors about the offers category could be exploited [13, 19, 20]. Finally, researchers could test the combination of different types of features to capture text and visual contents in other direct marketing domains.

Acknowledgements The authors would like to thank Daniele Ravì for helping in the implementation of the simulation studies. The authors would also like to thank Neodata Group for giving access to the mobile messaging dataset, and for helping in the implementation and testing of the proposed approach.

References

1. Alpaydin E (2004) Introduction to machine learning. MIT, Cambridge
2. Barnard K, Forsyth DA (2001) Learning the semantics of words and pictures. In: ICCV, Vancouver, 7–14 July 2001, pp 408–415
3. Battiato S, Farinella GM, Gallo G, Ravì D (2008) Scene categorization using bag of textons on spatial hierarchy. In: International conference on image processing (ICIP), San Diego, 12–15 October 2008
4. Battiato S, Farinella G, Giuffrida G, Tribulato G (2007) Data mining learning bootstrap through semantic thumbnail analysis. In: SPIE-IS&T 19th annual symposium electronic imaging science and technology 2007—multimedia content access: algorithms and systems, Orlando, 9–13 April 2007
5. Bergen JR, Julesz B (1983) Rapid discrimination of visual patterns. IEEE Trans Syst Man Cybern 13:857–863
6. Biederman I (1987) Recognition by components: a theory of human image interpretation. Psychol Rev 94:115–148
7. Biederman I, Mezzanotte R, Rabinowitz J (1982) Scene perception: detecting and judging objects undergoing relational violations. Cogn Psychol 14:143–177
8. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth and Brooks, Monterey
9. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm
10. Cleveland WS, Devlin SJ, Grosse E (1988) Regression by local fitting: methods, properties, and computational algorithms. J Econom 37(1):87–114
11. Comaniciu D, Ramesh V, Meer P (2003) Kernel-based object tracking. IEEE Trans Pattern Anal Mach Intell 25(5):564–575
12. Direct Marketing Association (2007) The power of direct marketing: ROI, sales, expenditures and employment in the U.S., 2006–2007 edn. Direct Marketing Association, Washington, DC
13. Florent P (2008) Universal and adapted vocabularies for generic visual categorization. IEEE Trans Pattern Anal Mach Intell 53(7):1243–1256
14. Hull D (1996) Stemming algorithms: a case study for detailed evaluation. J Am Soc Inf Sci 47: 70–84
15. Julesz B (1981) Textons, the elements of texture perception, and their interactions. Nature 290:91–97
16. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE conference on computer vision and pattern recognition, vol II. IEEE, Piscataway, pp 2169–2178
17. Li FF, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. In: CVPR '05: Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 2. IEEE Computer Society, Los Alamitos, pp 524–531
18. Lim JH (1999) Categorizing visual contents by matching visual "keywords". In: VISUAL, pp 367–374
19. Mairal J, Bach F, Ponce J, Sapiro G, Zisserman A (2008) Discriminative learned dictionaries for local image analysis. In: IEEE conference on computer vision and pattern recognition
20. Moosmann F, Triggs B, Jurie F (2007) Fast discriminative visual codebooks using randomized clustering forests. In: Schölkopf B, Platt J, Hoffman T (eds) Advances in neural information processing systems, vol 19. MIT, Cambridge, pp 985–992
21. Naccari F, Battiato S, Bruna A, Capra A, Castorina A (2005) Natural scene classification for color enhancement. IEEE Trans Consum Electron 5:234–239
22. Nash E (2000) Direct marketing. McGraw-Hill, New York
23. Netsize (2007) Convergence: everything is going mobile. The Netsize Guide 2007. Netsize, Levallois Perret
24. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42:145–175
25. Oren N (2002) Reexamining tf.idf based information retrieval with genetic programming. In: SAICSIT 2002, South African Institute for Computer Scientists and Information Technologists, Republic of South Africa, pp 224–234

26. Oza NC (2005) Online bagging and boosting. In: Systems, man and cybernetics, 2005 IEEE international conference on. IEEE, Piscataway, pp 2340–2345
27. Potter M (1975) Meaning in visual search. Science 187:965–966
28. Prinzie A, Van Den Poel D (2005) Constrained optimization of data-mining problems to improve model performance: a direct-marketing application. Expert Syst Appl 29(3):630–640
29. Renninger LW, Malik J (2004) When is scene recognition just texture recognition? Vis Res 44:2301–2311
30. Roberts M, Berger PD (1989) Direct marketing management. Prentice-Hall, New York
31. Schapire R (2001) The boosting approach to machine learning: an overview. Kluwer, Boston
32. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) New support vector algorithms. Neural Comput 12(5):1207–1245
33. Shawe-Taylor J, Cristianini N (2000) Support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
34. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. In: Proceedings of the international conference on computer vision, vol 2. IEEE, Piscataway, pp 1470–1477
35. Taylor P, Caley R, Black AW, King S (1999) Wagon, Edinburgh Speech Tools Library
36. Varma M, Zisserman A (2005) A statistical approach to texture classification from single images. Int J Comput Vis 62(1–2):61–81
37. Winn J, Criminisi A, Minka T (2005) Object categorization by learned universal visual dictionary. In: ICCV ’05: proceedings of the tenth IEEE international conference on computer vision. IEEE Computer Society, Washington, DC, pp 1800–1807
38. Yang J, Jiang YG, Hauptmann AG, Ngo CW (2007) Evaluating bag-of-visual-words representations in scene classification. In: MIR ’07: proceedings of the international workshop on multimedia information retrieval. ACM, New York, pp 197–206

Sebastiano Battiato was born in Catania, Italy, in 1972. He received the degree in Computer Science (summa cum laude) in 1995 and his Ph.D in Computer Science and Applied Mathematics in 1999. From 1999 to 2003 he has lead the “Imaging” team c/o STMicroelectronics in Catania. Since 2004 he works as a Researcher at Department of Mathematics and Computer Science of the University of Catania. His research interests include image enhancement and processing, image coding and camera imaging technology. He published more than 90 papers in international journals, conference proceedings and book chapters. He is co-inventor of about 15 international patents. He is reviewer for several international journals and he has been regularly a member of numerous international conference committees. He has participated in many international and national research projects. He is an Associate Editor of the SPIE Journal of Electronic Imaging (Specialty: digital photography and image compression). He is director of ICVSS (International Computer Vision Summer School). He is a Senior Member of the IEEE.

Giovanni Maria Farinella is currently contract researcher at Dipartimento di Matematica e Informatica, University of Catania, Italy (IPLAB research group). He is also associate member of the Computer Vision and Robotics Research Group at University of Cambridge since 2006. His research interests lie in the fields of computer vision, pattern recognition and machine learning. In 2004 he received his degree in Computer Science (egregia cum laude) from University of Catania. He was awarded a Ph.D. (Computer Vision) from the University of Catania in 2008. He has co-authored several papers in international journals and conferences proceedings. He also serves as reviewer numerous international journals and conferences. He is currently the co-director of the International Summer School on Computer Vision (ICVSS).

Giovanni Giuffrida is an assistant professor at University of Catania, Italy. He received a degree in Computer Science from the University of Pisa, Italy in 1988 (summa cum laude), a Master of Science in Computer Science from the University of Houston, Texas, in 1992, and a Ph.D. in Computer Science, from the University of California in Los Angeles (UCLA) in 2001. He has an extensive experience in both the industrial and academic world. He served as CTO and CEO in the industry and served as consultant for various organizations. His research interest is on optimizing content delivery on new media such as Internet, mobile phones, and digital tv. He published several papers on data mining and its applications. He is a member of ACM and IEEE.

Catarina Sismeiro is a senior lecturer at Imperial College Business School, Imperial College London. She received her Ph.D. in Marketing from the University of California, Los Angeles, and her Licenciatura in Management from the University of Porto, Portugal. Before joining Imperial College Catarina had been and assistant professor at Marshall School of Business, University of Southern California. Her primary research interests include studying pharmaceutical markets, modeling consumer behavior in interactive environments, and modeling spatial dependencies. Other areas of interest are decision theory, econometric methods, and the use of image and text features to predict the effectiveness of marketing communications tools. Catarina's work has appeared in innumerous marketing and management science conferences. Her research has also been published in the *Journal of Marketing Research*, *Management Science*, *Marketing Letters*, *Journal of Interactive Marketing*, and *International Journal of Research in Marketing*. She received the 2003 Paul Green Award and was the finalist of the 2007 and 2008 O'Dell Awards. Catarina was also a 2007 Marketing Science Institute Young Scholar, and she received the D. Antonia Adelaide Ferreira award and the ADMES/MARKTEST award for scientific excellence. Catarina is currently on the editorial boards of the *Marketing Science* journal and the *International Journal of Research in Marketing*.

Giuseppe Tribulato was born in Messina, Italy, in 1979. He received the degree in Computer Science (summa cum laude) in 2004 and his Ph.D in Computer Science in 2008. From 2005 he has lead the research team at Neodata Group. His research interests include data mining techniques, recommendation systems and customer targeting.