Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Quality-Based Learning for Web Data Classification

Ou Wu, Ruiguang Hu, Xue Mao, Weiming Hu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences.
{wuou, rghu, xmao, wmhu}@nlpr.ia.ac.cn

Abstract

The types of web data vary in terms of information quantity and quality. For example, some pages contain numerous texts, whereas some others contain few texts; some web videos are in high resolution, whereas some other web videos are in low resolution. As a consequence, the quality of extracted features from different web data may also vary greatly. Existing learning algorithms on web data classification usually ignore the variations of information quality or quantity. In this paper, the information quantity and quality of web data are described by quality-related factors such as text length and image quantity, and a new learning method is proposed to train classifiers based on quality-related factors. The method divides training data into subsets according to the clustering results of quality-related factors and then trains classifiers by using a multi-task learning strategy for each subset. Experimental results indicate that the quality-related factors are useful in web data classification, and the proposed method outperforms conventional algorithms that do not consider information quantity and quality.

Introduction

The Internet has become indispensable in people's daily life. Therefore, the need to classify and manage web data increases. Although much achievement has been made in previous web data classification, and encouraging results have been obtained, the complexity of web data is not well considered in existing studies. Web pages are designed by humans. The designers and information sources of different pages are distinct, which results in that the types of web data, including texts, images, and videos, are complex. The types of web data vary in two aspects:

- *Information quantity is usually distinct*. Take web pages as an example. Some pages contain many images, whereas some pages contain few images. Some pages contain numerous texts, whereas some other pages contain few texts. This phoneme still exists for images. Some web images have many text descriptions, whereas some other web images have limited text descriptions. Figure 1 shows three web pages with different proportion of texts and images. In Fig. 1(a), the page contains a number of images and few texts; in Fig. 1(c), the page contains few images but plentiful texts. Figure 2 shows three examples of web images with different lengthes of text descriptions.
- *Information quality is usually distinct*. The quality of web images and videos is greatly affected by factors such as the performance of capture devices and the environment. As many web images and videos are produced by low-quality devices, they are with low resolutions or distorted colors. Figure 3 illustrates how videos with similar contents differ in quality (e.g., resolution and color distortion). It is very likely that the Fig. 3(a) video is obtained by a low-quality camera.

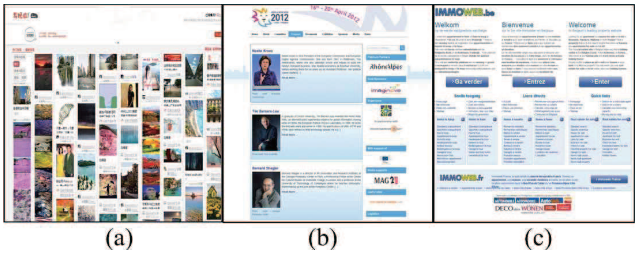

Figure 1: Three web pages with different proportions of images and texts.

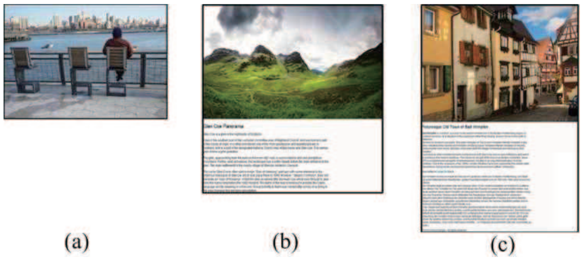

Figure 2: Three images with different lengths of text descriptions.

Variations in information quantity and quality of web data result in the variations of the quality of extracted features. Intuitively, features with different quality levels should make unequal contributions in the final classification. For example, in Fig. 1, image features (or text features) should make distinct contributions in the classification of Fig. 1(a) and Fig. 1(c) pages. Likewise, text features should make distinct contributions when classifying the three images in Fig. 2. Therefore, information quantity and quality of web data should ideally be considered during classifier training and

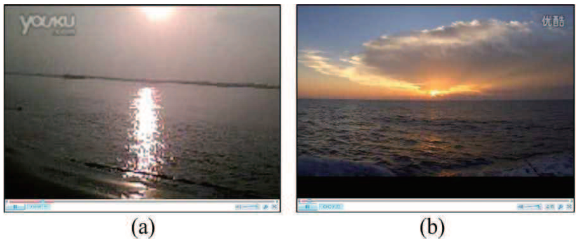

Figure 3: Two web videos with different visual quality.

classification. To our knowledge, little headway has been made along this way in web data classification. Considering that information quantity can also be viewed as a quality measure for web information, the factors related to both information quantity and quality are called *quality-related factors*. Some typical quality-related factors are the text length of a web document, the image count in a web page, and visual quality of an web image or video. Indeed, information quality has been explored in web information retrieval. Bendersky *et al.* (2011) proposed a quality-biased web document ranking algorithm based on the notion that the quality of real web documents is usually not identical. Kumar *et al.* (2011) took visual quality as an attribute in image search.

This paper proposes a new learning method which integrates the quality-related factors of web data both in the model training and in the classification of new data. The integration of quality-related factors in classification has been investigated in biometrics (Nandakumar *et al.*, 2008)(Poh and Kittler, 2012). However, obvious differences exist between this work and the quality-based fusion in biometrics: (1) this work focuses on learning classifiers, whereas quality-based fusion focuses on fusion and assumes that classifiers are given; (2) the proposed method can be used for single-modal data, whereas quality-based fusion is designed only for multi-modal data. Our work is original in the following aspects:

- To our knowledge, this is the first time that information quantity and quality are considered in web data classification. Quality-related factors[1], such as text length, illumination, and video quality, are used to describe the information quantity and quality. They are used in both learning and classification in this work.
- A new quality-based learning method is proposed. The method divides the training set into subsets according to the clusters of quality-related factors. A multi-task learning approach is introduced to learn classifiers in each subset. Both hard and soft clustering strategies are investigated and two concrete algorithms (LQHC and LQSC) are obtained.

Related Work

Two studies are closely related to this work. One is the quality-based fusion in biometrics. The other involves classification of web data such as web pages and web images.

Recent studies on multi-modal biometrics give attention to the quality-based fusion because the quality of biometric data is usually negatively affected by factors such as environment, noise, and devices (Kittler *et al.*, 2007). Poh and Kittler (2012) proposed a unified framework for quality-based fusion of multi-modal biometrics. The framework only pursues dynamic fusion strategies while quality-based learning in this work pursues both dynamic fusion strategies and classifier parameters.

Classifying web pages and its containing elements (e.g., texts, image, and videos on the web) can be used for constructing web directories, improving quality of search results, and filterign web content (Xu *et al.*, 2007). A recent survey can be found in (Qi and Davison, 2009). All existing studies ignore the quality (and quantity) of information used for feature extraction and successive classification. However, like biometric data, the quality also affects the feature extraction and subsequent classification for web data.

Methodologies

To begin with, an intuitive learning algorithm is introduced which gives more weights to the features with higher quality. Then, its disadvantages are discussed. Finally, motivated by this simple algorithm, a new learning method is proposed.

An intuitive algorithm

Web data classification usually employs multi-modal features. Different modality features usually have different quality levels. For simplicity, assume that two modalities are present. Let X_a be the feature space for the first modality and X_v be the feature space for the second modality. Then for the ith sample, x_{ai} and x_{vi} are the features for the two modalities, respectively. Each sample is associated with two quality-related factors for the two modalities. The two quality-related factors for the ith sample are represented by $q_{ai} \in [0, 1]$ and $q_{vi} \in [0, 1]$, respectively. A higher value of a quality-related factor indicates a higher quality of its corresponding features. Let Y be the output space whose elements are '-1' or '1'.

Intuitively, the higher the quality of the features from one modality, the larger the weight of the features in the final classifier. Assuming that the classifier is linear, then the classifier (f) that integrates features and quality-related factors can be represented by the following equation:

$$f(x_i) = \frac{q_{ai}}{s_i}(w_a^T x_{ai} + b_a) + \frac{q_{vi}}{s_i}(w_v^T x_{vi} + b_v) \tag{1}$$

where w_a, b_a, w_v, and b_v are the classifier parameters for the two types of features; $s_i = q_{ai} + q_{vi}$.

To learn the classifier parameters w_a, b_a, w_v, and b_v, the framework of the support vector machine (SVM) is used. First, Eq. (1) is re-written as

$$\begin{aligned} f(x_i) &= \tfrac{q_{ai}}{s_i}(w_a^T x_{ai} + b_a) + \tfrac{q_{vi}}{s_i}(w_v^T x_{vi} + b_v) \\ &= [w_a^T, w_v^T] \begin{bmatrix} \frac{q_{ai}}{s_i} x_{ai} \\ \frac{q_{vi}}{s_i} x_{vi} \end{bmatrix} + (1 - \tfrac{q_{vi}}{s_i}) b_a + \tfrac{q_{vi}}{s_i} b_v \\ &= [w_a^T, w_v^T, w_b] \begin{bmatrix} \frac{q_{ai}}{s_i} x_{ai} \\ \frac{q_{vi}}{s_i} x_{vi} \\ \frac{q_{vi}}{s_i} \end{bmatrix} + b_a \end{aligned} \tag{2}$$

[1] It should be noted that we are not required to provide the quality-factors manually. Instead, all the factors can be automatically obtained in a similar way to feature extraction.

where $w_b = b_v - b_a$. If we denote

$$\bar{w} = [w_a^T, w_v^T, w_b]^T, \bar{x}_i = [\tfrac{q_{ai}}{s_i} x_{ai}^T, \tfrac{q_{vi}}{s_i} x_{vi}^T, \tfrac{q_{vi}}{s_i}]^T, \bar{b} = b_a \tag{3}$$

Then, the objective function of the SVM here is defined as

$$\begin{aligned} &\min_{\bar{w}, \bar{b}, \xi_i} \tfrac{1}{2}||\bar{w}|| + C \sum_{i=1}^{N} \xi_i \\ &s.t. \quad \forall i : y_i(\bar{w}^T \bar{x}_i + \bar{b}) \geq 1 - \xi_i, \qquad \xi_i > 0 \end{aligned} \tag{4}$$

where C controls the model complexity, and ξ_i is the slack factor. (4) can be solved with similar techniques for the SVM. Once $\bar{w}$ and $\bar{b}$ are obtained, the new feature vector for a test sample is calculated by using Eq. (3) using its raw features (x_i) and quality-related factors(q_{ai} and q_{vi}). The label is then achieved by using Eq. (2).

In this intuitive algorithm, the (normalized) quality-related factors are taken as weights of the features. Therefore, the above learning with quality weight algorithm is called LQW. In practice, LQW suffers from three problems:

- LQW linearly combines quality-related factors and features. However, the relationship between quality-related factors and features may be not exactly linear. In this case, the linear combination is inaccurate.
- LQW considers that only one quality-related factor exists for each modality. However, the quality-related factors for each modality may be more than one.
- LQW deals only with multi-modal features. However, some factors affect the feature quality in some cases with single-modality features.

The proposed method

Equation (1) can be re-written as

$$\begin{aligned} f(x_i) &= [\tfrac{q_{ai}}{s_i} w_a^T, \tfrac{q_{vi}}{s_i} w_v^T] \begin{bmatrix} x_{ai} \\ x_{vi} \end{bmatrix} + (\tfrac{q_{ai}}{s_i} b_a + \tfrac{q_{vi}}{s_i} b_v) \\ &= w_{q_i}^T x_i + b_{q_i} \end{aligned} \tag{5}$$

where

$$w_{q_i} = [\frac{q_{ai}}{s_i} w_a^T, \frac{q_{vi}}{s_i} w_v^T]^T \quad \text{and} \quad b_{q_i} = (\frac{q_{ai}}{s_i} b_a + \frac{q_{vi}}{s_i} b_v) \tag{6}$$

As shown in the above equations, for any two samples, if their quality-related factors are similar, then their corresponding classifiers (parameterized by w_{q_i} and b_{q_i}) are also similar. Motivated by this observation, we propose a new method which learns a specific classifier for samples with similar quality-related factors. *First*, the quality-related factors of the training samples are clustered. *Then* the obtained clusters are used to divide the training samples into training subsets. This step ensures that samples within a training subset have similar quality-related factors. *Finally*, samples in each training subset are used to train a classifier.

Assuming that M clusters of quality-related factors are obtained, for the mth cluster's corresponding training subset (called the mth training subset), its classifier is $f_m(x) = w_m^T x + b_m$. Let X_m and Y_m be the mth training subset. Then w_m and b_m are obtained by solving the following equation

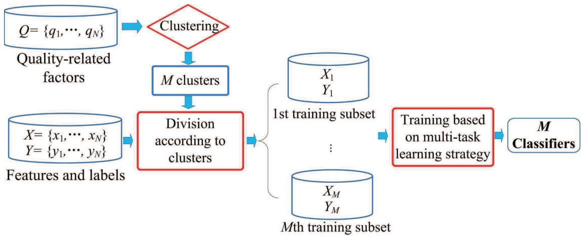

Figure 4: The overall of the proposed method.

$$\min_{w_m, b_m} \sum_{j=1}^{N_m} L(y_{mj}, w_m^T x_{mj} + b_m) + \gamma R(w_m), x_{mj} \in X_m \tag{7}$$

where $R(w_m)$ is the regularization term.

Considering that the learning tasks for w_m and b_m for each training subset are similar and correlated, a multi-task learning strategy is used to achieve all the classifiers for each training subset. Learning multiple related tasks simultaneously has been shown to improve significantly the performance relative to learning each task independently (Liu *et al.*, 2009). The overall of this method is shown in Fig. 4.

Let $W = [w_1, \cdots, w_M]$ and $B = [b_1, \cdots, b_M]^T$. The optimization function of the multi-task feature learning for W and B is

$$\min_{(W,B)} \{ \sum_{m=1}^{M} \sum_{j=1}^{N_m} L(y_{mj}, w_m^T x_{mj} + b_m) + \gamma \left\| W \right\|_{2,1} \} \tag{8}$$

where N_m is the number of training samples in the mth cluster; $y_{mj} \in Y_m$ and $x_{mj} \in X_m$. (8) can be solved with the multi-task feature learning technique.

The above approach is based on the "hard" assignment of a quality-related factor to clusters. Nevertheless, a "hard" assignment does not consider cluster ambiguity (Liu *et al.*, 2011). To this end, a soft clustering algorithm, the Gaussian mixture model (GMM), is used.

Assume the mth cluster of quality-related factors is modeled by a Gaussian distribution with parameters π_m, μ_m, and Σ_m. An iteration strategy can be used to maximize this function and to obtain the parameters. For an input sample associated with the quality-related factor q_i, the probability that the sample belongs to the mth cluster is

$$p(m|q_i) = \frac{p(m, q_i)}{p(q_i)} = \frac{\pi_m \mathrm{N}(q_i|\mu_m, \Sigma_m)}{\sum\limits_{m=1}^{M} \pi_m \mathrm{N}(q_i|\mu_m, \Sigma_m)} \tag{9}$$

For each training sample (or a test sample), we obtain a vector of conditional probabilities as follows:

$$P_i = (p(1|q_i), \cdots, p(M|q_i))^T \tag{10}$$

As a consequence, the predicted label of a sample is

$$f(x_i) = \sum_{m=1}^{M} P_i(m)(w_m^T x_i + b_m) = P_i{}^T (W^T x_i + B) \tag{11}$$

The multi-task feature learning with the soft clustering is

$$\begin{aligned} &\min_{(W,B)} \{ \sum_{i=1}^{N} L(y_i, \sum_{m=1}^{M} P_i(m) w_m^T x_i + \sum_{m=1}^{M} P_i(m) b_m) + \gamma \left\| W \right\|_{2,1} \} \\ &= \min_{W,B} \{ \sum_{i=1}^{N} L[y_i, P_i^T (W^T x_i + B)] + \gamma \left\| W \right\|_{2,1} \} \end{aligned} \tag{12}$$

When the square loss is used for (12), we define

$$\Omega(W,B) = \sum_{i=1}^{N} \left(y_i - {P_i}^T (W^T x_i + B)\right)^2 + \gamma \left\| W \right\|_{2,1} \tag{13}$$

$\Omega(W,B)$ is decomposed as follows:

$$\begin{aligned}\Omega(W,B) = \textstyle\sum_{i=1}^{N} \{&y_i^2 - 2y_i P_i^T B + P_i^T B P_i^T B - 2y_i P_i^T W^T \\ &+P_i^T W^T x_i P_i^T W^T x_i + 2P_i^T B P_i^T W^T x_i\} + \gamma \left\| W \right\|_{2,1} \}\end{aligned} \tag{14}$$

Note that $P_i^T W^T x_i$ is a value, then

$$\begin{aligned}&\tfrac{\partial \Omega(W,B)}{\partial W} = \textstyle\sum_{i=1}^{N} [\tfrac{\partial x_i^T W P_i P_i^T W^T x_i}{\partial W} + 2\tfrac{\partial (P_i^T B - y_i) P_i^T W^T x_i}{\partial W}] \\ &+\gamma \tfrac{\partial \left\| W \right\|_{2,1}}{\partial W}\end{aligned} \tag{15}$$

We also have

$$\tfrac{\partial x_i^T W P_i P_i^T W^T x_i}{\partial W} = \tfrac{\partial tr(W P_i P_i^T W^T x_i x_i^T)}{\partial W} = 2x_i x_i^T W P_i P_i^T \tag{16}$$

Note that

$$\frac{\partial {P_i}^T W^T x_i}{\partial W} = \frac{\partial tr(P_i {x_i}^T W)}{\partial W} = x_i {P_i}^T \tag{17}$$

(15) becomes

$$\frac{\partial \Omega(W,B)}{\partial W} = \sum_{i=1}^{N} \{2x_i x_i^T W P_i P_i^T + 2({P_i}^T B - y_i) x_i {P_i}^T\} + 2\gamma D W \tag{18}$$

D is a diagonal matrix and its ith diagonal element is[2] $2||W^{(i)}||_2{}^{-1}$. Similarly,

$$\tfrac{\partial P_i^T B P_i^T B}{\partial B} = \tfrac{\partial tr(B P_i^T B P_i^T)}{\partial B} = 2 P_i B^T P_i \tag{19}$$

We obtain

$$\tfrac{\partial \Omega(W,B)}{\partial B} = 2(\textstyle\sum_{i=1}^{N} y_i P_i - \sum_{i=1}^{N} P_i {P_i}^T W^T x_i - \sum_{i=1}^{N} P_i B^T P_i) \tag{20}$$

For W, let the values of (18) be zero. We obtain

$$\gamma D W + \sum_{i=1}^{N} x_i x_i^T W P_i P_i^T = \sum_{i=1}^{N} (y_i - {P_i}^T B) x_i {P_i}^T \tag{21}$$

For B, let the values of (20) be zero, we obtain

$$\sum_{i=1}^{N} y_i P_i - \sum_{i=1}^{N} P_i {P_i}^T W^T x_i - \sum_{i=1}^{N} P_i B^T P_i = 0 \tag{22}$$

Note that $PB^T P = PP^T B$. Equation (22) becomes

$$B = (\sum_{i=1}^{N} P_i {P_i}^T)^{-1} (\sum_{i=1}^{N} y_i P_i - \sum_{i=1}^{N} P_i {P_i}^T W^T x_i) \tag{23}$$

Thus, a heuristic solution for W and B is proposed. In each iteration, the values of W and B are updated using

$$\begin{aligned}&\gamma W^{(t+1)} + (D^{(t)})^{-1} \textstyle\sum_{i=1}^{N} x_i x_i^T W^{(t+1)} P_i P_i^T \\ &= (D^{(t)})^{-1} \textstyle\sum_{i=1}^{N} (y_i - {P_i}^T B^{(t)}) x_i {P_i}^T\end{aligned} \tag{24}$$

$$B^{(t+1)} = (\sum_{i=1}^{N} P_i {P_i}^T)^{-1} (\sum_{i=1}^{N} y_i P_i - \sum_{i=1}^{N} P_i {P_i}^T (W^{(t+1)})^T x_i) \tag{25}$$

Once W is pursued, features are selected according to W. The classifiers are then trained with the selected features. The classifier of the mth training subset is learned by solving

$$\begin{aligned}&\min_{{w'}_m, b'_m, \xi_i} \tfrac{1}{2} ||{w'}_m|| + C \textstyle\sum_{i=1}^{N} P_i(m) \xi_i \\ &s.t. \quad \forall i : y_i [{w'}_m^T {x'}_i + {b'}_m] \geq 1 - \xi_i, \qquad \xi_i > 0\end{aligned} \tag{26}$$

where w'_m and b'_m are the classifier parameters of the mth training subset; x'_i is the new feature of x_i based on the selected features. Given that the above algorithm is based on the clustering of quality-related factors, the algorithm is called LQHC when the clustering is hard and LQSC when the clustering is soft. Compared with LQW, both LQHC and LQSC have three advantages:

- The quality-related factors are implicitly used and are not assumed to be linear with the features. In LQW, the factors are assumed to be linear with their corresponding features.
- The number of quality-related factors for each modality is not limited. In LQW, the number must be one.
- The features are not required to be multi-modal. In LQW, the features should be multi-modal.

The algorithmic steps of LQSC are summarized in Algorithm **1**. The The algorithmic steps of LQHC are similar and omitted due to lack of space.

Algorithm 1 Learning (and testing) based on the soft clustering for quality-related factors (LQSC)

```
Input: Training data $(X, Y)$ and associated quality-related factors
    $Q$; a test sample $x_t$ and its quality-related factor $q_t$, $M$, $T$.
Initialize: $W^{(0)}$, $B^{(0)}$.
Steps:
1. Cluster quality-related factors $Q$ into $M$ groups using GMM;
2. Calculate $P_i$ for each training sample by using Eq. (10);
3. Learn the feature weights $W$ by iteratively updating $W$ and $B$
   by using Eqs. (24) and (25) until the maximum number of iterations ($T$) is attained or the iteration is converged;
4. Select features according to $W$;
5. Learn $M$ classifiers with selected features for each training subset by solving (26);
6. Calculate the probability vector $P(q_t)$ by using Eq. (10);
7. Calculate the new feature vector ($x'_t$) of $x_t$ based on $W$;
8. Classify $x'_t$ by using the $M$ classifiers, $P(q_t)$, and Eq. (11);
Output: The GMM of all the $M$ clusters of quality-related factors,
    the $M$ classifiers, and the predicted label of $x_t$.
```

Experiments

Experimental setup

Two common-usedly classification algorithms, namely SVM and random forest (RF) (Breiman, 2001), are used

[2] When $W^{(i)} = 0$, the value of d_{ii} cannot be calculated. Nevertheless, it is observed from Eq. (24) that only D^{-1} is required.

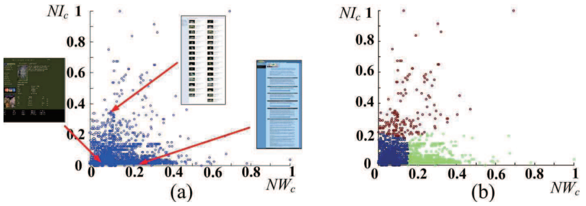

Figure 5: (a) The distribution of NI_C and NW_C on the cannabis web page data set. (b) The clustering results.

as the baseline competing methods. Another intuitive algorithm, which directly takes quality-related factors as additional features, is also compared. This algorithm directly combines the features and quality-related factors as a new feature vector for each sample, so it is called **direct concatenation**. The radial basis kernel is chosen for both SVM and LQW. The parameters C and g are searched via five-cross validation in $\{0.1, 1, 10, 50, 100\}$ and $\{0.001, 0.01, 0.1, 1, 10\}$, respectively. For the SVM used in LQHC and LQSC, the parameters are searched with the same settings. For RF, only the number of trees in $\{10, 50, 100, 200, 300\}$ is changed, and other parameters are default. Specifically, the parameter γ in LQHC and LQSC is searched in $\{0.0001, 0.001, 0.01, 0.1, 1\}$. For the direct concatenation algorithm, the SVM is used. The maximum number of iterations used in LQSC is set to 20. Three measures, namely, precision, recall, and $F1$, are used.

Results on cannabis web page recognition

Illicit cannabis web pages pose a negative influence on users, especially teenagers (Wang *et al.*, 2011). The data set consisting of 4427 normal and cannabis web pages in (Wang *et al.*, 2011) is used. Given a web page, let I_c be its image count and W_c be its word count. They are normalized as follows: $NI_c = min(I_c/80, 1)$ and $NW_c = min(W_c/8000, 1)$.

The distribution of NI_c and NW_c of the collected pages is shown in Fig. 5(a). Some pages contain more than 2000 words, whereas some pages contain no more than 10 words. Some pages contain more than 50 images, whereas some pages contain no image. Three typical pages are also shown in Fig. 5(a). The parameters NI_c and NW_c are taken as the quality-related factors[3] of each page. The clustering results with K-means for NI_c and NW_c are shown in Fig. 5(b). In Fig. 5(b), the pages are divided into three clusters, namely, image dominant (the top cluster), text dominant (the right cluster), and mixture of images and texts. We have also observed that the clusters do not have clear margins. Therefore, using a soft clustering strategy appears more reasonable than that using a hard strategy.

The document frequency method is used for text features. A total of 100 words are used. Therefore, the text features for each page are a 100-dimensional vector. A page usually contains more than one image. The image features are extracted as follows. First, the standard scale-invariant feature transform (Lowe, 2004) is used for local patch description, and the bag of word model (Csurka *et al.*, 2004) is used to construct the histogram for each image. Second, all histograms are clustered into K subsets. All the images of each page are allocated into K clusters, and the normalized histogram of the numbers of images in all the K clusters is taken as the feature vector. In the experiments, K is set to 50. Therefore, the image features of each page consist of a 50-dimensional vector. The text and image features of each page are concatenated, and a 150-dimensional feature vector is obtained.

Table 1 shows the classification results of the seven competing algorithms. In both LQHC and LQSC, the number of clusters (M) is set as 3. All the four learning algorithms using quality-related factors (Direct concatenation, LQW, LQHC, and LQSC) achieve better results compared with the other three algorithms which are based on features alone. The $F1$ value of LQSC is about 4.36% higher than that of the SVM which does not utilize quality-related factors. To test the robustness of LQHC and LQSC, we perform both algorithms under different numbers of clusters (M). The recognition results of LQHC and LQSC with the increasing of M in terms of the $F1$ values. When $M = 1$, the $F1$ values of both algorithms are equal. The value is 0.9051 which is higher than that of SVM. The reason is that when $M = 1$, the two algorithms are identical to the approach of feature selection via $l_{2,1}$-norm and SVM. When $M \geq 3$, both algorithms achieve significant better $F1$ values than the other algorithms. When M equals 6, the $F1$ values of LQHC and LQSC are 0.9511 and 0.9649, respectively. The partial reason for the performance improvement is that with the increase of M, the quality-related factors in each training subset vary slightly and become more similar with each other. Further more, although the numbers of samples in each training subset become smaller leading that the corresponding classifiers may be insufficiently learned, the multi-task learning used here alleviates this problem by transferring knowledge among training subsets.

Table 1: The results on the cannabis web page recognition.

	Precision	Recall	$F1$
SVM (only features)	0.9323	0.8563	0.8926
RF (only features)	0.9291	0.8580	0.8921
Wang *et al.* (2011) (only features)	0.9211	0.8933	0.9070
Direct concatenation	0.9195	0.9001	0.9097
LQW	0.9590	0.8908	0.9234
LQHC ($M = 3$)	0.9676	0.8887	0.9265
LQSC ($M = 3$)	0.9781	0.8983	**0.9365**

Results on pornographic image recognition

Recently, pornographic image recognition has attracted much attention in both academic research and industrial application. Most existing algorithms rely on the skin features of images. Therefore, skin detection is a key step and severs as the basis in many previous algorithms. However, the illumination of web images is very complexity. Figure 6 shows normal images from the Internet. The top three images feature the same person. However, the skin colors change under different illumination conditions. The bottom three images are captured by Phone or PC cameras and have low-

[3] It should be noted that some other factors such as the number of hyperlinks and the image sizes can be also taken as quality-related factors. These factors will be considered in our future work.

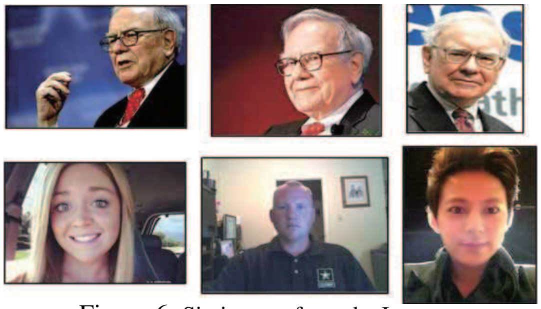

Figure 6: Six images from the Internet.

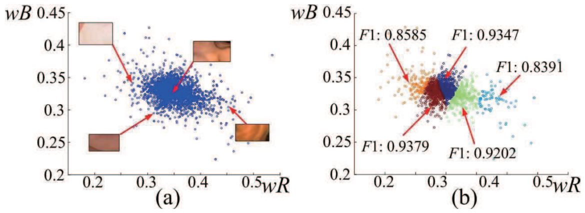

Figure 7: (a) The distribution of the quality-related factors of the pornographic image set and some skin patches. (b) The clusters of the quality-related factors and the $F1$ values.

quality illumination conditions. Considering that skin detection plays a crucial role in existing studies, we evaluate the quality of detected skin pixels and then apply the quality into succeeding model training and classification.

Assessing directly the quality of extracted skin pixels for pornographic image classification is difficult. Note that the quality of extracted skin pixels is most affected by illumination (Hu *et al.*, 2007). Therefore, we adopt an alternative strategy. First, we estimate the illumination of each image. We then cluster the illumination and sort images with similar illumination conditions into the same cluster. Consequently, the quality levels of detected skin pixels of the images in the same training subset may be similar. The algorithm proposed by Weijer *et al.* (2007) is applied to estimate the illumination of an input image. The algorithm outputs the illumination color with two quantities (wR, wB) which are taken as the quality-related factors for an image.

The image data introduced in (Zuo *et al.*, 2010) is applied. The distribution of the estimated illumination is shown in Fig. 7. The images in some areas have bad illumination conditions. Figure 7(a) also shows the skin patches of some sample images. The colors of skins with different illumination conditions vary significantly.

To explore the relationship between the classification performance and illumination, we divide the data set according to the estimated illumination. The corresponding data subset for each cluster is random split into two equal parts. One part is used for training and the other is used for testing. The random split is repeated 10 times. A SVM classifier is used and the average classification results are recorded. Finally, the $F1$ values of the different clusters' corresponding data subsets are obtained. Figure 7(b) shows the clustering of quality-related factors and the $F1$ results. The clusters with worse illumination have lower $F1$ values.

The estimated illumination in this data set cannot be directly used as weights as it has two components. Therefore, the LQW algorithm cannot be used on this data set. The skin detection and feature extraction adapt the methods used by Zuo et al. (Zuo *et al.*, 2010). Table 2 shows the classification results of the five competing methods. In both LQHC and LQSC, the number of clusters (M) is set as 3. For LQHC and LQSC, the number of clusters is set to 3. All the learning algorithms using quality-related factors (Direct concatenation, LQHC, and LQSC) still achieve better results than the others do. The $F1$ value of the LQSC method is about 4.22% higher than that of the SVM without considering information quality. We then perform both LQHC and LQSC under different numbers of clusters. Similar observations to those from cannabis page recognition are obtained. Both LQHC and LQSC show good performances.

Table 2: The results on the pornographic image recognition.

	Precision	Recall	$F1$
SVM (only features)	0.9097	0.8920	0.9008
RF (Zuo *et al.*, 2010) (only features)	0.9196	0.9018	0.9106
Direct concatenation	0.9243	0.9161	0.9202
LQHC ($M = 3$)	0.9325	0.9144	0.9234
LQSC ($M = 3$)	0.9524	0.9339	**0.9430**

Discussion

Several observations are obtained from the above experiments. (1) The quality-related factors do improve the classification performance for web data with distinct information quantity or quality. In both experiments, the algorithms (Direct concatenation, LQW, LQHC, and LQSC) which integrate the quality-related factors outperform the others without integration. (2) LQHC and LQSC achieve better results than Direct concatenation and LQW which simply take quality-related factors as additional features and weights, repsectively. (3) LQSC outperforms LQHC on both sets. As shown in Figs. 5(b) and 7(b), there are no clear boundary between clusters. Consequently, a soft clustering strategy appears more reasonable than a hard strategy.

Conclusions

This paper has investigated the classification problems for web data with unequal information quantity or quality. A new learning method has been proposed which divides the whole training data into different subsets according to the clustering of their associated quality-related factors, and then learns models for each subset. Using different clustering strategies, two learning algorithms have been obtained, namely, LQHC and LQSC. The results of two experiments further validate the effectiveness of our proposed method. In addition, LQSC, which employs a soft clustering strategy, is better than LQHC which employs a hard strategy.

Acknowledgment

We would like to thank Prof. Shiming Xiang for the very helpful suggestions. This work is partly supported by NSFC (Grant No. 61379098, 61003115, 61103056) and Baidu research fund.

References

Breiman, L., Random forests, *Machine Learning*, 45, pp. 5-32, 2001.

Csurka, G., Dance, C., Fan, L., Williamowski, J., and Bray, C., Visual categorization with bags of keypoints, *In Proceedings of ECCV International Workshop on Statistical Learning in Computer Vision (SLCV)*, pp. 1-22, 2004.

Hu, W., Wu, O., Chen, Z., Fu, Z., and Maybank, S., Recognition of Pornographic Web Pages by Classifying Texts and Images, *IEEE Trans. Pattern Anal. Mach. Intell. (TPMAI)* 29, 6, pp. 1019-1034, 2007.

Hu, W., Zuo, H., Wu, O., Chen, Y., Zhang, Z., and Suter, D., Recognition of adult images, videos, and web page bags, *ACM Trans. Multimedia Comput. Commun. Appl.* 7S, 1, Article 28, 2011.

Kalva, P., Enembreck, F., and Koerich, A., Web Image Classification Based on the Fusion of Image and Text Classifiers, *In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, pp. 561-568, 2007.

Kittler, J., Poh, N., Fatukasi, O., Messer, K., Kryszczuk, K., Richiardi, J., and Drygajlo, A., Quality dependent fusion of intramodal and multimodal biometric experts, *Proc. SPIE 6539, Biometric Technology for Human Identification IV*, 653903, 2007.

Liu, J., Ji, S., and Ye, J., Multi-Task Feature Learning Via Efficient l2,1-Norm Minimization, I*n Proceedings of Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 339-348, 2009.

Liu, L., Wang, L., and Liu, X., In Defense of Soft-assignment Coding, *In Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 2486-2493, 2011.

Lowe, D., Distinctive image features from scale-invariant keypoints, *In Proceedings of International Journal of Computer Vision (IJCV)*, vol. 60, pp. 91-110, 2004.

Nandakumar, K., Chen, Y., Dass, S. C., and Jain, A. K., *Likelihood Ratio-Based Biometric Score Fusion*, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 2, pp. 342-347, 2008.

Nie, F., Huang, H., Cai, X., and Ding, C., Efficient and Robust Feature Selection via Joint l2, 1-Norms Minimization, *In Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp. 1813-1821, 2010.

Poh, N. and Kittler, J., A Unified Framework for Biometric Expert Fusion Incorporating Quality Measures, *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34 no. 1, pp. 3-18, 2012.

Qi, X. and Davison, B. D., Web page classification: Features and algorithms, *ACM Comput. Surv.* 41, 2, Article 12, 2009.

Song, Y., Zhou, D., Huang, J., Councill, I. G., Zha, H., and Giles, C. L., Boosting the Feature Space: Text Classification for Unstructured Data on the Web, *In Proceedings of the Sixth International Conference on Data Mining (ICDM)*, pp. 1064-1069, 2006.

Wang, Y., Xie, N., Hu, W., and Yang, J., Multi-Modal Multiple-Instance Learning with the Application to the Cannabis Webpage Recognition, *In Proceedings of Asian Conference on Pattern Recognition (ACPR)*, pp. 105 - 109, 2011.

Wang, Z., Zhao, M., Song, Y., Kumar, S., and Li, B., *YouTubeCat: Learning to Categorize Wild Web Videos*, *In Proceedings of IEEE International Conference on Computer Vision (CVPR)*, pp. 879-886, 2010.

Weijer, J. van de, Gevers, T., and Gijsenig, A., Edge-Based Color Constancy, *IEEE Trans. Img. Proc.* 16, 9, pp. 2207-2214, 2007.

Xu, Z., King, I., and Lyu, M-R., Web page classification with heterogeneous data fusion, *In Proceedings of the 16th International Conference on World Wide Web (WWW)*, pp. 1171-1172, 2007.

Zhou, J., Chen, J., and Ye, J., Clustered Multi-Task Learning Via Alternating Structure Optimization, *In Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pp.702-710, 2011.

Zuo, H., Hu, W., and Wu, O., Patch-based skin color detection and its application to pornography image filtering, *In Proceedings of International Conference on World Wide Web (WWW)*, pp. 1227-1228, 2010.

Dekel, O., and Shamir, O., Vox populi: Collecting high-quality labels from a crowd. *Annual Conference on Learning Theory (COLT)*, 2009.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K., Describable Visual Attributes for Face Verification and Image Search, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol. 33, No. 10, pp. 1962-1977, 2011.