Bayesian logistic regression for classification of tabular data*

Kropotov D., Vetrov D., Wolf L., Hassner T.
dmitry.kropotov@gmail.com, vetrovd@yandex.ru, wolf@cs.tau.ac.il, hassner@openu.ac.il
Moscow, Russia, Dorodnicyn Computing Centre of RAS, Moscow State University;
Tel-Aviv, Israel, Tel-Aviv University, The Open University of Israel

We extend the Relevance Vector Machine (RVM) framework to handle cases of table-structured data, i. e. when each object is represented by a table of features rather than by a single vector. This is achieved by coupling the regularization coefficients of rows and columns of features. We present two variants of this new gridRVM framework, based on the way in which the regularization coefficients of the rows and columns are combined. Appropriate variational optimization algorithms are derived for inference within this framework. The consequent reduction in the number of parameters from the product of the table's dimensions to the sum of its dimensions allows for better performance in the face of small training sets, resulting in improved resistance to overfitting problems, as well as providing better interpretation of results. These properties are demonstrated on a synthetic data-set as well as on a modern and challenging visual identification benchmark.

In classical machine learning theory, a training set consists of a number of objects (precedents), each represented as a vector of features. This is not, however, always an optimal representation. In some cases, a tabular representation is more convenient. Objects are then described by a number of features that form a table rather than a single vector.

A natural example of such case arises in a region/descriptor-based framework for image analysis. Within this framework, an image is split into several regions (blocks) and a set of descriptors is then computed for each region. Then, we may associate each feature with the pair region/descriptor and form a tabular view of a single image. Note that often the number of features extracted from the image exceeds the number of images in the whole training set, resulting in increased risk of overfitting.

Another example is related to the use of radial basis functions (RBF) in classification algorithms. Traditionally, RBF depends only on the distance $\rho(\boldsymbol{x}, \boldsymbol{y}_m)$ between the object $\boldsymbol{x}$ and some predefined point $\boldsymbol{y}_m$ in the space of features $\mathbb{R}^d$, i. e. $\varphi_m(\boldsymbol{x}) = f\bigl(\rho(\boldsymbol{x}, \boldsymbol{y}_m)\bigr)$, $m = 1, \ldots, M$. Each object is described by a vector of M RBF values. Gaussian RBFs $\varphi_m(\boldsymbol{x}) = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{y}_m\|^2)$ are a popular “rule-of-thumb” choice in many classification algorithms, e. g. in logistic regression. The obvious drawback of Gaussian RBFs is their low discriminative ability in the presence of numerous noisy features. To deal with noisy features one might consider basis functions consisting of a single feature $\varphi_{mj}(\boldsymbol{x}) = f\bigl(|x(j) - y_m(j)|\bigr)$. Although it is possible to represent an object $\boldsymbol{x}$ as a vector $(\varphi_{11}(\boldsymbol{x}), \ldots, \varphi_{M,d}(\boldsymbol{x}))$, it could be more natural to form a table of M columns and d rows.

The tabular representation of data provides new options in analyzing feature sets. In particular, we may search for relevant columns and rows instead of searching for relevant features. Besides, we will show that in some cases tabular data classifier has better generalization properties compared to analogous feature vector classifier.

GridRVM models

Consider a two-class classification problem with tabular data. Let $(X, \boldsymbol{t}) = \{\boldsymbol{x}_n, t_n\}_{n=1}^N$ be the training set where $t_n \in \{-1, 1\}$ are class labels and each object $\boldsymbol{x}_n$ is represented as a table of generalized features $\bigl(\varphi_{ij}(\boldsymbol{x}_n)\bigr)_{i,j=1}^{M_1,M_2}$. Note that we will also use one-index notation $(\varphi_k(\boldsymbol{x}_n))_{k=1}^M$, $M = M_1M_2$ when we need to treat the description of the object as a vector and denote $\Phi = \{\varphi_k(\boldsymbol{x}_n)\}_{k,n=1}^{M,N}$. Define the following probabilistic model (p-gridRVM):

$$p(\boldsymbol{t}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta} \,|\, X) = p(\boldsymbol{t} \,|\, X, \boldsymbol{w}) p(\boldsymbol{w} \,|\, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}),$$

$$p(\boldsymbol{t} \,|\, X, \boldsymbol{w}) = \prod_{n=1}^N \sigma\bigl(t_n \boldsymbol{w}^{\intercal} \boldsymbol{\varphi}(\boldsymbol{x}_n)\bigr),$$

$$p(\boldsymbol{w} \,|\, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\prod\limits_{i,j=1}^{M_1,M_2} \sqrt{\alpha_i \beta_j}}{\sqrt{2\pi}^{M_1 M_2}} \exp\Biggl(-\frac{1}{2} \sum_{i,j=1}^{M_1,M_2} \alpha_i \beta_j w_{ij}^2\Biggr), \quad (1)$$

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{M_1} \mathcal{G}(\alpha_i \,|\, a_0, b_0), \; p(\boldsymbol{\beta}) = \prod_{j=1}^{M_2} \mathcal{G}(\beta_j \,|\, c_0, d_0),$$

where $\sigma(y) = 1/\bigl(1 + \exp(-y)\bigr)$ is a logistic function, $\mathcal{G}(\alpha_i \,|\, a_0, b_0)$ stands for gamma distribution over α_i with parameters a_0, b_0 and all $\alpha_i, \beta_j \geqslant 0$. The p-gridRVM model differs from the conventional RVM model only in (1), where instead of individual regularization coefficient α_{ij} for each weight w_{ij} we assign independent regularization coefficients to each row and column of the tabular presentation. The regularization coefficient for the weight w_{ij} is a product of α_i and β_j. Alternatively, we may consider the sum, i. e.

$$p(\boldsymbol{w} \,|\, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\prod\limits_{i,j=1}^{M_1,M_2} \sqrt{\alpha_i + \beta_j}}{\sqrt{2\pi}^{M_1 M_2}} \exp\Biggl(-\frac{1}{2} \sum_{i,j=1}^{M_1,M_2} (\alpha_i + \beta_j) w_{ij}^2\Biggr).$$

The work is supported by the Russian Foundation for Basic Research (project No. 09-01-92474) and Ministry of Science, Culture and Sport, State of Israel (project No. 3-5797).

We refer to this model as s-gridRVM. Note that in both introduced models the number of regularization coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is M_1+M_2 while the number of weights is M_1M_2. However, they have one important distinction. In s-gridRVM large value of α_i means that all features from the i^{th} row have regularization coefficients at least as large as α_i, while in p-gridRVM large α_i does not necessarily imply large values of the regularization coefficient for a weight w_{ij} since the coefficient β_j may have a small value. Thus we may expect a different behavior from these models.

Variational learning in gridRVM models

In a classification problem we wish to calculate

$$p(t_{\text{new}}\,|\,\boldsymbol{x}_{\text{new}},\boldsymbol{t},X)=\int p(t_{\text{new}}\,|\,\boldsymbol{x}_{\text{new}},\boldsymbol{w})\times p(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}\,|\,\boldsymbol{t},X)\,d\boldsymbol{w}\,d\boldsymbol{\alpha}\,d\boldsymbol{\beta} \quad (2)$$

for any new object $\boldsymbol{x}_{\text{new}}$. For models p- and s-gridRVM this integration is intractable and hence some approximation scheme is needed. Here we use the variational approach [1], which has been successfully applied for the conventional RVM model in [3], and try to find a variational approximation $q(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta})$ to the true posterior $p(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}\,|\,\boldsymbol{t},X)$ in the following family of factorized distributions:

$$q(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta})=q_{\boldsymbol{w}}(\boldsymbol{w})q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})q_{\boldsymbol{\beta}}(\boldsymbol{\beta}).$$

Then (2) can be reduced to integration over the factorized distribution q:

$$p(t_{\text{new}}|\boldsymbol{x}_{\text{new}},\boldsymbol{t},X)\simeq\int p(t_{\text{new}}|\boldsymbol{x}_{\text{new}},\boldsymbol{w})q_{\boldsymbol{w}}(\boldsymbol{w})d\boldsymbol{w}. \quad (3)$$

Using the Jaakkola-Jordan inequality [1] for the likelihood function $p(\boldsymbol{t}\,|\,X,\boldsymbol{w})$, we obtain:

$$p(\boldsymbol{t}\,|\,X,\boldsymbol{w})\geqslant F(\boldsymbol{t},X,\boldsymbol{w},\boldsymbol{\xi})=\prod_{n=1}^{N}\sigma(\xi_n)\exp\Big(\frac{z_n-\xi_n}{2}-\lambda(\xi_n)(z_n^2-\xi_n^2)\Big),$$

where $\lambda(\xi)=\tanh(\xi/2)/(4\xi)$, $z_n=t_n\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\varphi}(\boldsymbol{x}_n)$. This bound is tight for $\xi_n=z_n$. Then it can be shown that

$$\log p(\boldsymbol{t}\,|\,X)\geqslant\int\log\frac{F(\boldsymbol{t},X,\boldsymbol{w},\boldsymbol{\xi})p(\boldsymbol{w}\,|\,\boldsymbol{\alpha},\boldsymbol{\beta})p(\boldsymbol{\alpha})p(\boldsymbol{\beta})}{q_{\boldsymbol{w}}(\boldsymbol{w})q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})q_{\boldsymbol{\beta}}(\boldsymbol{\beta})}\times q_{\boldsymbol{w}}(\boldsymbol{w})\,q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})\,q_{\boldsymbol{\beta}}(\boldsymbol{\beta})\,d\boldsymbol{w}\,d\boldsymbol{\alpha}\,d\boldsymbol{\beta}. \quad (4)$$

From the theory of variational inference [1], it follows that maximization of the criterion function (4) w. r. t. distributions $q_{\boldsymbol{w}}(\boldsymbol{w})$, $q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$, $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ and variational parameters $\boldsymbol{\xi}$ leads to the following result:

$$q_{\boldsymbol{w}}(\boldsymbol{w})=\mathcal{N}(\boldsymbol{w}\,|\,\boldsymbol{\mu},\Sigma), \quad (5)$$
$$q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})=\prod_{i=1}^{M_1}\mathcal{G}(\alpha_i\,|\,a_i,b_i),\quad q_{\boldsymbol{\beta}}(\boldsymbol{\beta})=\prod_{j=1}^{M_2}\mathcal{G}(\beta_j\,|\,c_j,d_j),$$

where each distribution is iteratively updated with all others fixed by the following formulae:

$$\Sigma=\left(\operatorname{diag}(\mathsf{E}_{\boldsymbol{\alpha}}\alpha_i\mathsf{E}_{\boldsymbol{\beta}}\beta_j)+2\Phi^{\mathsf{T}}\Lambda\Phi\right)^{-1},$$
$$\Lambda=\operatorname{diag}\big(\lambda(\xi_n)\big),\quad \boldsymbol{\mu}=\frac{1}{2}\Sigma\Phi^{\mathsf{T}}\boldsymbol{t},$$
$$a_i=a_0+\frac{M_2}{2},\quad b_i=b_0+\frac{1}{2}\sum_{j=1}^{M_2}\mathsf{E}_{\boldsymbol{\beta}}\beta_j\mathsf{E}_{\boldsymbol{w}}w_{ij}^2,$$
$$c_j=c_0+\frac{M_1}{2},\quad d_j=d_0+\frac{1}{2}\sum_{i=1}^{M_1}\mathsf{E}_{\boldsymbol{\alpha}}\alpha_i\mathsf{E}_{\boldsymbol{w}}w_{ij}^2,$$
$$\xi_n^2=\boldsymbol{\varphi}^{\mathsf{T}}(\boldsymbol{x}_n)\mathsf{E}_{\boldsymbol{w}}\boldsymbol{w}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\varphi}(\boldsymbol{x}_n).$$

The necessary statistics are calculated as follows:

$$\begin{array}{ll}\mathsf{E}_{\boldsymbol{w}}\boldsymbol{w}=\boldsymbol{\mu}, & \mathsf{E}_{\boldsymbol{\alpha}}\log\alpha_i=\Psi(a_i)-\log b_i,\\ \mathsf{E}_{\boldsymbol{w}}w_{ij}^2=\Sigma_{ij,ij}+\mu_{ij}^2, & \mathsf{E}_{\boldsymbol{\beta}}\beta_j=\frac{c_i}{d_i},\\ \mathsf{E}_{\boldsymbol{\alpha}}\alpha_i=\frac{a_i}{b_i}, & \mathsf{E}_{\boldsymbol{\beta}}\log\beta_j=\Psi(c_j)-\log d_j,\end{array} \quad (6)$$

where $\Psi(a)=\frac{d}{da}\log\Gamma(a)$ — digamma function.

For learning in s-gridRVM model we propose a new variational bound:

$$\log(x+y)\geqslant\log(\eta+\zeta)+\frac{\eta\big(\log(x)-\log(\eta)\big)+\zeta\big(\log(y)-\log(\zeta)\big)}{\eta+\zeta}, \quad (7)$$

where η and ζ are variational parameters. This bound is tight when $x/y=\eta/\zeta$ and illustrated in Fig. 1. The inequality (7) leads to the following lower bound on $\log p(\boldsymbol{w}\,|\,\boldsymbol{\alpha},\boldsymbol{\beta})$:

$$\log p(\boldsymbol{w}\,|\,\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{2}\sum_{i,j=1}^{M_1,M_2}[\log(\alpha_i+\beta_j)-(\alpha_i+\beta_j)w_{ij}^2]-\frac{M_1M_2}{2}\log 2\pi\geqslant\log G(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\zeta})=$$
$$\frac{1}{2}\sum_{i,j=1}^{M_1,M_2}\left[\log(\eta_{ij}+\zeta_{ij})+\frac{\eta_{ij}\big(\log(\alpha_i)-\log(\eta_{ij})\big)}{\eta_{ij}+\zeta_{ij}}+\frac{\zeta_{ij}\big(\log(\beta_j)-\log(\zeta_{ij})\big)}{\eta_{ij}+\zeta_{ij}}-(\alpha_i+\beta_j)w_{ij}^2\right]-\frac{M_1M_2}{2}\log 2\pi.$$

This bound is tight, e. g. if $\eta_{ij}=\alpha_i$ and $\zeta_{ij}=\beta_j$. Maximization of the criterion function

$$\log p(\boldsymbol{t}\,|\,X)\geqslant\int\log\frac{F(\boldsymbol{t},X,\boldsymbol{w},\boldsymbol{\xi})G(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\zeta})p(\boldsymbol{\alpha})p(\boldsymbol{\beta})}{q_{\boldsymbol{w}}(\boldsymbol{w})q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})q_{\boldsymbol{\beta}}(\boldsymbol{\beta})}\times q_{\boldsymbol{w}}(\boldsymbol{w})q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})q_{\boldsymbol{\beta}}(\boldsymbol{\beta})d\boldsymbol{w}d\boldsymbol{\alpha}d\boldsymbol{\beta}$$

w. r. t. distributions $q_{\boldsymbol{w}}(\boldsymbol{w})$, $q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$, $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ and variational parameters $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, $\boldsymbol{\zeta}$ leads to (5), where

$$\Sigma=\big(\operatorname{diag}(\mathsf{E}_{\boldsymbol{\alpha}}\alpha_i+\mathsf{E}_{\boldsymbol{\beta}}\beta_j)+2\Phi^{\mathsf{T}}\Lambda\Phi\big)^{-1},$$
$$\Lambda=\operatorname{diag}\big(\lambda(\xi_n)\big),\ \boldsymbol{\mu}=\frac{1}{2}\Sigma\Phi^{\mathsf{T}}\boldsymbol{t},$$

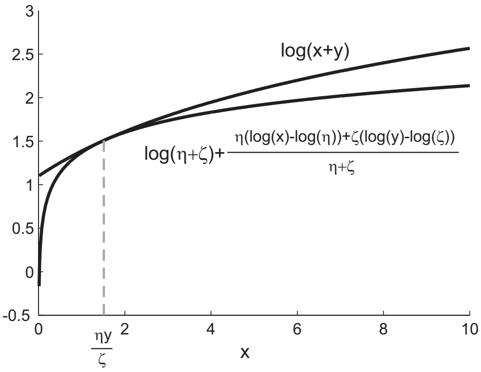


Fig. 1. One-dimensional projection of the bound (7) for parameters $y = 3,\ \eta = 2,\ \zeta = 4$.

$$a_i = a_0 + \frac{1}{2}\sum_{j=1}^{M_2}\frac{\eta_{ij}}{\eta_{ij}+\zeta_{ij}}, \qquad b_i = b_0 + \frac{1}{2}\sum_{j=1}^{M_2}\mathsf{E}_{\boldsymbol{w}} w_{ij}^2,$$

$$c_j = c_0 + \frac{1}{2}\sum_{i=1}^{M_1}\frac{\zeta_{ij}}{\eta_{ij}+\zeta_{ij}}, \qquad d_j = d_0 + \frac{1}{2}\sum_{i=1}^{M_1}\mathsf{E}_{\boldsymbol{w}} w_{ij}^2,$$

$$\eta_{ij} = \exp(\mathsf{E}_{\boldsymbol{\alpha}}\log\alpha_i),\ \ \zeta_{ij} = \exp(\mathsf{E}_{\boldsymbol{\beta}}\log\beta_j),$$

$$\xi_n^2 = \boldsymbol{\varphi}^{\mathsf{T}}(\boldsymbol{x}_n)\mathsf{E}_{\boldsymbol{w}}\boldsymbol{w}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\varphi}(\boldsymbol{x}_n).$$

The necessary statistics are still calculated using (6).

Now return to decision making scheme (3). Using (5) the integral (3) can be rewritten as

$$\int\sigma\big(t_{\text{new}}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\varphi}(\boldsymbol{x}_{\text{new}})\big)\mathcal{N}(\boldsymbol{w}\,|\,\boldsymbol{\mu},\Sigma)d\boldsymbol{w} =$$

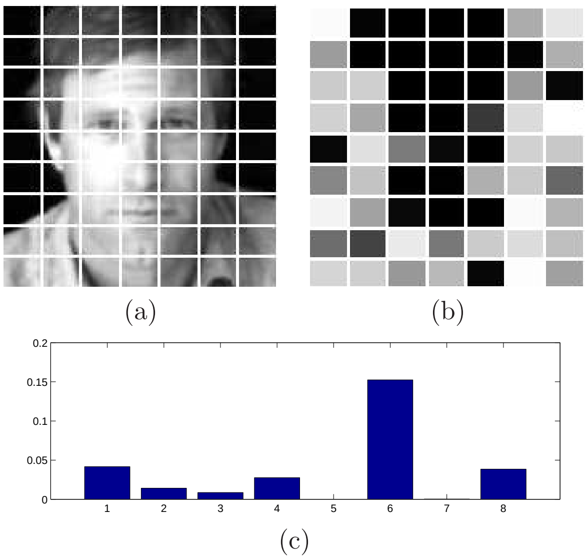
$$\int\sigma(z)\mathcal{N}\big(z\,|\,t_{\text{new}}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\varphi}(\boldsymbol{x}_{\text{new}}),\boldsymbol{\varphi}^{\mathsf{T}}(\boldsymbol{x}_{\text{new}})\Sigma\boldsymbol{\varphi}(\boldsymbol{x}_{\text{new}})\big)dz.$$

The last integral is one-dimensional and can be easily calculated using the Monte Carlo technique. The useful analytical approximation for this integral is proposed in [5]:

$$\int\sigma(z)\mathcal{N}(z\,|\,m,s^2)dz \simeq \sigma\left(m\Big/\sqrt{1+\tfrac{\pi s^2}{8}}\right).$$

Experiments

First consider an artificial classification dataset[1] taken from [4] (see Fig. 3c). This is a 2-class problem with 200 objects in the training set and 5000 objects in the test set. The feature space is two-dimensional and the data are generated from a specified distribution with Bayesian error rate 19%. The optimal discriminative surface is non-linear. In the experiment we add up to 30 normally distributed noisy features and investigate the behaviour of the conventional variational RVM [3], p-gridRVM and s-gridRVM with 3 types of basis functions. In the first case we take initial features, i. e. $\varphi_j(\boldsymbol{y}) = y(j)$ (total d features). This corresponds to a linear separating hyperplane. In the second case we take Gaussian RBFs of the form $\varphi_j(\boldsymbol{y}) = \exp(-\delta\|\boldsymbol{y}-\boldsymbol{x}_j\|^2)$, where $\boldsymbol{x}_j$ are training objects (total N features). In the third case we take separate RBFs calculated for each dimension, i. e. $\varphi_{ij}(\boldsymbol{y}) = \exp\Big(-\delta\big(y(i)-x_j(i)\big)^2\Big)$ (total Nd features). In the first two cases we have a standard vector representation of objects, $M_2 = 1$ for both gridRVMs and hence gridRVMs are very similar to standard RVM here. In the last case we may treat objects' representation both as a matrix of size $N\times d$ for gridRVMs and as a vector of length Nd for RVM. The experimental results (error rates) are shown in figure 3 (a: initial features, b: standard RBFs, d: RBFs calculated for each dimension, color legend: light grey — RVM, black — p-gridRVM, dark grey — s-gridRVM, dotted line stands for train error, solid line — test error). In all cases $\delta = 5.55$. In the first case we have more than 27% error rate for all three methods because linear hyperplane is inadequate for this nonlinear data. For the second case all methods show similar performance and quickly overfit with the addition of noisy features. However, the overfit speed for gridRVM methods is less than for RVM. In the last case, where the tabular representation of data is appropriate, gridRVM methods show stable performance resulting in 22–23% of errors even for 30 noisy features while RVM definitely overfits starting from several noisy features. The number of the relevant basis functions for the last type of basis functions is shown on fig. 3e. We can see that the s-gridRVM model gives less sparse solution compared to p-gridRVM.


(a) (b)

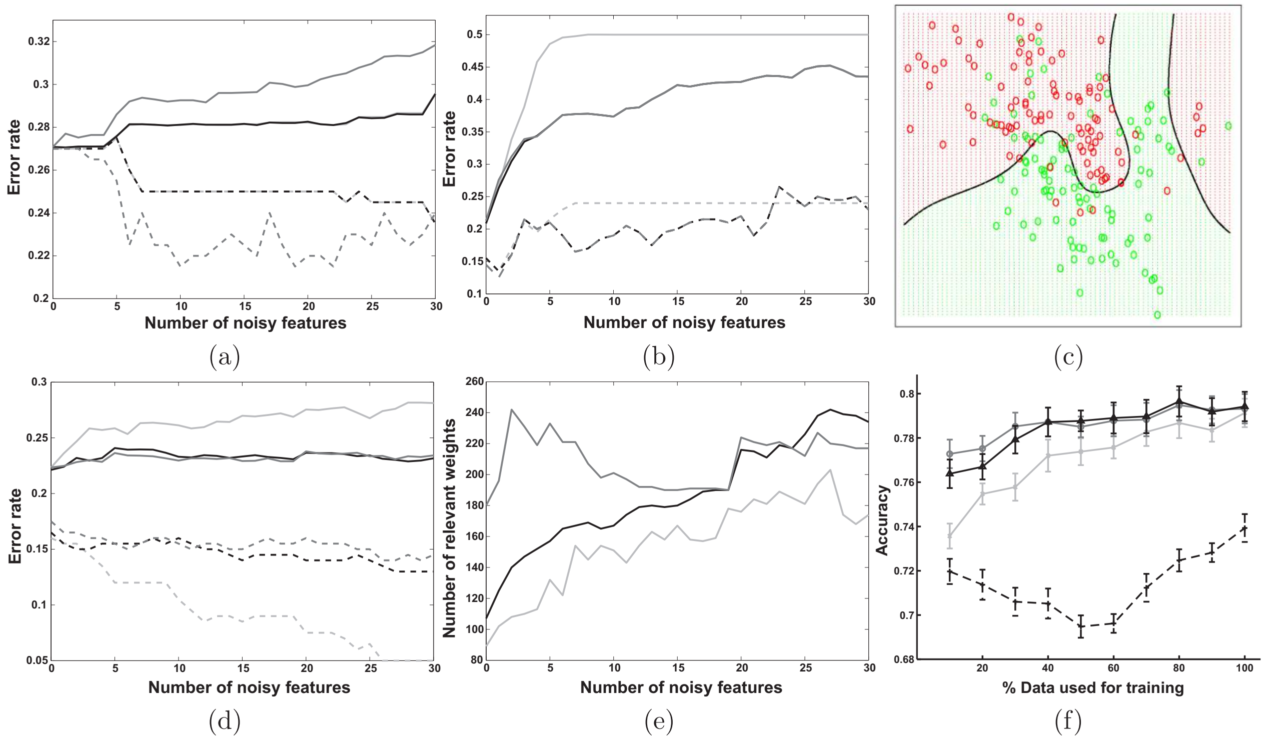
(c)

Fig. 2. Experimental results for LFW dataset. See the text for details.

We also test gridRVM approach on the Labeled Faces in the Wild (LFW) pair-matching benchmark[2].

[1] http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/mixture.example.data

[2] http://vis-www.cs.umass.edu/lfw/



(a) (b) (c)

(d) (e) (f)

Fig. 3. Experimental results. Please see the text for details.

The LFW data set provides a set of facial images which were automatically harvested from news websites and thus present faces under challenging, unconstrained viewing conditions. The goal of the benchmark is to determine, given a pair of images from the collection, whether the two images match (portray the same subject) or not.

We represent the images in the following way. Each face image was subdivided into 63 non-overlapping blocks of 23×18 pixels centered on the face (see Fig. 2a) and for each block a set of values was calculated using Local Binary Patterns approach [6]. We used four different LBPs and thus four different feature vectors for each block. Each pair of images to be compared is represented by one table of similarity values. The rows of the tables correspond to similarities values, calculated using particular LBP type and particular distance function (we used two variants – Euclidian and Hellinger distance) and the columns correspond to the 63 facial blocks.

We report our results in Fig. 3f where the pair-matching performance of s-gridRVM (dark grey) and p-gridRVM (black) is compared against two baseline methods — RVM (light grey) and linear SVM (dotted black). As can be seen, the gridRVM methods show a clear advantage over both baseline methods. This is particularly true when only a small amount of training data is available. Although this advantage diminishes as more training is made available, both grid methods remain superior. Note that the results improve the ones reported in [2], where the same features were used for the whole image and the reported (best at that time) accuracy was 78.47%. P-gridRVM and s-gridRVM showed 79.34% and 79.42% of correct answers respectively.

GridRVM approach allows us to analyze relevant rows and columns in object's tabular representation. In the context of face images this corresponds to relevant blocks and relevant descriptors (LBPs + distance types). Fig. 2b shows the block relevance (the darker the more informative) and Fig. 2c shows the relevance of descriptors (inverse regularization coefficient).

References

[1] *Jaakkola T., Jordan M.* Bayesian parameter estimation through variational methods // Statistics and Computing. — 2000. — Vol. 10. — Pp. 25–37.

[2] *Wolf L., Hassner T., Taigman Y.* Descriptor based methods in the wild // ECCV workshop on Faces in Real-Life Images: Detection, Alignment, and Recognition. — 2008.

[3] *Bishop C., Tipping M.* Variational relevance vector machine // Proc. of the 16-th Conf. on Uncertainty in Artificial Intelligence. — 2000. — Pp. 46–53.

[4] *Friedman J., Hastie T., Tibshirani R.* The elements of statistical learning. — New York: Springer, 2001.

[5] *MacKay D.* Bayesian interpolation // Neural Computation. — 1992. — Vol. 4, No. 3. — Pp. 415–447.

[6] *Ahonen T., Hadid A., Pietikäinen M.* Face description with local binary patterns: application to face recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2006. — Vol. 28, No. 12. — Pp. 2037–2041.