Int. J. Mach. Learn. & Cyber.
DOI 10.1007/s13042-013-0156-6

ORIGINAL ARTICLE

Sparse group LASSO based uncertain feature selection

Zongxia Xie · Yong Xu

Received: 5 January 2013 / Accepted: 19 February 2013

Abstract Uncertain data management and mining is becoming a hot topic in recent years. However, little attention has been paid to uncertain feature selection so far. In this paper, we introduce the sparse group least absolution shrinkage and selection operator (LASSO) technique to construct a feature selection algorithm for uncertain data. Each uncertain feature is represented with a probability density function. We take each feature as a group of values. Through analysis of the current four sparse feature selection methods, LASSO, elastic net, group LASSO and sparse group LASSO, the sparse group LASSO is introduced to select feature selection from uncertain data. The proposed algorithm can select not only the features between groups, but also the sub-features in groups. As the trained weights of feature groups are sparse, the groups of features with weight zero are removed. Experiments on nine UCI datasets show that feature selection for uncertain data can reduce the number of features and sub-features at the same time. Moreover it can produce comparable accuracy with all features.

Keywords Uncertain data · Feature selection · Sparse group LASSO

1 Introduction

Uncertainty widely exists in real-world data due to noise, quantization or limitation of measurement [1]. Large scale of uncertain data bring a lot of challenges to the domains of databases and data analysis [4, 18, 33]. How to manage and analyze uncertain information is becoming a hot topic in recent years. Some workshops and special issues have been organized to discuss the related problems [3, 7, 16].

In fact, uncertainty is the essential property of data in practical applications because of instrument errors, modeling errors and sampling errors, repeated measurements and so on [1, 31]. Some studies have focused on how to model with uncertain data effectively in recent years. Support vector machines [5, 29], associative classifiers [25], Naive Bayes classifiers [26] and decision tree [34] were extended to construct models for uncertain data. These researches show that if we adequately take the uncertainty into account, the classification performances are improved.

Feature selection is one of the most important research fields in pattern recognition and machine learning [12, 28]. It can remove irrelevant and/or redundant features and bring some advantages at the same time [13, 30]. For examples, prediction performance can be improved and algorithms can be faster run. As we know that a collection of algorithms have been developed for feature selection. However, to the best of our knowledge, no work has been focused on uncertain feature selection so far except [9]. Unlike the classical modeling tasks, the feature values of uncertain data are usually described with a probability density function, instead of a single value. So the feature selection algorithms for classical tasks are not applicable to evaluate the quality of uncertain features. These algorithms can not be directly used in the new case. In this paper, we focus on feature selection algorithms for uncertain data. we expect that the above advantages of feature selection will emerge in uncertain data.

Each feature of uncertain data is not a single value, but a set of values which form a probability distribution. This

Z. Xie (✉) · Y. Xu
Bio-Computing Research Center, Shenzhen Graduate School,
Harbin Institute of Technology, Shenzhen 518055, China
e-mail: caddiexie@hotmail.com

Y. Xu
e-mail: laterfall2@yahoo.com.cn

Published online: 10 March 2013

causes the dimension of data increases sharply. Some researchers started to study feature selection algorithms for uncertain data to solve this problem. Doquire and Verleysen introduced mutual information to select features for uncertain data in [9]. The mutual information is first computed between each feature of the training set. Then the features are ranked according to this score. They show that their method is effective to select relevant features. However, this technique does not consider the grouping of features.

Sparsity techniques are shown to be very effective in feature selection, classification and regression learning, even ensemble pruning [10]. In this paper, we propose a sparse group (least absolution shrinkage and selection operator) LASSO method to conduct feature selection for uncertain data. This method is an extension of LASSO and group LASSO [11, 24], which can yield a solution that achieves the within- and between- group sparsity simultaneously. This idea is right suitable for feature selection of uncertain data. Each feature described with a set of values is taken as a group. We are interested in identifying important feature groups as well as important values within the selected features.

The rest of the paper is structured as follows. In Sect. 2, descriptions and properties of uncertain data are given. Section 3 presents the proposed algorithm based on sparse group LASSO for uncertain data. Experimental results are shown and analyzed in Sect. 4. Section 5 concludes this paper and gives some future work.

2 Basic definitions

One way to model the uncertainty is to represent each feature as a probability density function (PDF). In this case, a dataset $(X, Y) = \{(x_i, y_i) | i = 1, 2, \ldots, l\}, x_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$ presents l samples with d features. x_{ij} stands for the j feature of the i sample, which is a PDF. In this paper, we consider binary classification problem. Therefore, $y_i \in \{-1, +1\}$. We assume that $[a_{ij}, b_{ij}]$ is a bounded interval of x_{ij} and $\overline{x_{ij}}$ is the mean of the j feature of the i sample. $[a_{ij}, b_{ij}]$ is set to

$$w \cdot |A_j|. \tag{1}$$

A_j denotes the width of the range for the feature j and w is a parameter, which controls the amplitude of uncertainty.

In order to implement numerically, we sample s points within $[a_{ij}, b_{ij}]$ to approximate this PDF by a discrete distribution. Here, s is the sampling number. The larger s is, the richer information there is. However, it take much time to processing large number of sample points.

Most physical measures involve random noise, which follows the Gaussian distribution. While digitization of the measured values introduces quantization noise, which satisfied the uniform distribution [34]. In this paper, we mainly consider Gaussian and uniform distributions. For Gaussian noise model, we use $\frac{1}{4}(b_{ij} - a_{ij})$ as the standard deviation. For the uniform noise model, the PDF is $(b_{ij} - a_{ij})^{-1}$. Figure 1 shows data samples with uncertainty. We see that data around the centers form ellipse for Gaussian uncertainty or a rectangle for uniform uncertainty.

How to measure the distance or similarity is a key issue in analyzing uncertain data. In order to take the points on PDF as a whole, we introduce the histogram intersection kernel (HIK) to measure the similarity of two uncertain data. HIK commonly employed in computer vision. It is proposed to compare distributions (histograms) of low level features in images, such as bag of visual words and spatial pyramids [17].

Histogram intersection is represented as follows. We use X_1 and X_2 as the histograms of two images. Assume that X_1 and X_2 have the same size (N pixels) and both histograms consist of m bins, and the l_bth bin for $l_b = 1, \ldots, m$ is

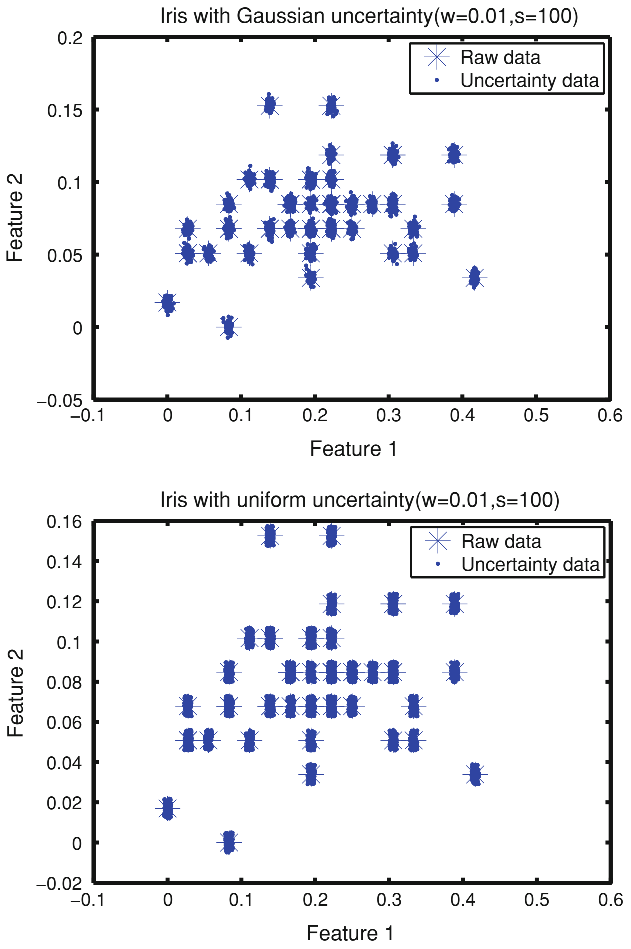

Fig. 1 Two features of Iris data with different amplitude ($w = 0.01$) of different uncertainty (Gaussian distribution and uniform distribution). Sampling number s is 100

denoted with x_{1l_b} and x_{2l_b} respectively. We have $\sum_{l_b=1}^{m} x_{1l_b} = N$ and $\sum_{l_b=1}^{m} x_{2l_b} = N$. Then histogram intersection is

$$K_{\mathrm{HIK}}(x_1, x_2) = \sum_{l_b=1}^{m} \min\{x_{1l_b}, x_{2l_b}\}. \tag{2}$$

Therefore, HIK can be easily used to compare uncertain samples with only one PDF feature. What's more, it can also extend to apply on multiple features of PDF with the properties of kernels. As we known, the sum of positive definite kernels are again a positive definite kernel [8]. With HIK, we can first compute the similarity of the samples with each feature, and then compute the sum of all the features.

3 Algorithm design

In this section, we first introduce several sparse feature selection methods and analyze their properties. Then sparse group LASSO is chosen to design an algorithm for selecting features from uncertain data.

3.1 Sparse techniques for feature selection

In the statistical community, the LASSO is a shrinkage and variable selection method , which is a penalized least square method and imposes an l_1-penalty on the coefficients [32]. LASSO can be formulated as the following optimization problem:

$$\begin{aligned} &\min_w \tfrac{1}{2}||Xw - Y||_2^2 + \lambda||w||_1 \\ &s.t.\ \ \lambda > 0 \end{aligned} \tag{3}$$

where the first term of the optimization problem is the least square loss and λ is the regularization parameter. Due to the nature of the l_1-penalty, the LASSO continuously shrinks the coefficients w toward zero and λ controls the degree of sparsity. The bigger λ is, the more number of zero in w is.

Because the underlying representations of many real-world processes are often sparse. During the past decade, sparse learning via l_1 regularization and its various extensions have received increasing attention in many areas including machine learning and statistics. In particular, sparse feature selection has been extensively investigated on both optimization algorithms and statistical properties. The above method has been extended to many new algorithms in order to solve different problems. We will introduce three sparse methods: elastic net, group LASSO, and sparse group LASSO as follows.

Zou and Hastie [36] proposed a new regularization technique called elastic net. The elastic net can not only do automatic feature selection but also select groups of correlated variables. However, the LASSO tends to select only one feature from the group features among which the pairwise correlations are very high. The elastic net can be represented by

$$\begin{aligned} &\min_w \tfrac{1}{2}||Xw - Y||_2^2 + \lambda_1||w||_1 + \lambda_2||w||_2^2 \\ &s.t.\ \ \lambda_1 > 0, \lambda_2 > 0. \end{aligned} \tag{4}$$

If there are group structures in features of data, sparse modeling has been explored for group feature selection. The group LASSO [35] utilizes the group information of the features and yields a solution with group sparsity. It introduces an l_2-regularization method for each group, which ultimately yields a group-wisely sparse model. The optimization problem of group LASSO is

$$\begin{aligned} &\min_w \tfrac{1}{2}||Xw - Y||_2^2 + \lambda\sum_{i=1}^{g}||W_{G_i}||_2 \\ &s.t.\ \ \lambda > 0 \end{aligned} \tag{5}$$

where w is divided into g non-overlapping groups $\mathrm{w}_{G1}, \mathrm{w}_{G2}, \ldots, \mathrm{w}_{Gg}$. The group LASSO gives a sparse set of groups. If it includes a group in the model then all coefficients in the group will be nonzero. If the size of each group is 1, the group LASSO degenerates to the regular LASSO solution. Huang and Zhang [14] showed that the group LASSO is superior to the standard LASSO under the strong group sparsity and certain other conditions, including a group sparse eigenvalue condition.

The sparse group LASSO is a collection of regularization methods, combing the LASSO and the group LASSO [11]. The sparse group LASSO penalty yields a solution that achieves the within- and between- group sparsity simultaneously. It enables to encourage sparsity at the level of both features and groups simultaneously. The optimization problem is formulated as

$$\begin{aligned} &\min_w \tfrac{1}{2}||Xw - Y||_2^2 + \lambda_1||w||_1 + \lambda_2\sum_{i=1}^{g}||W_{G_i}||_2 \\ &s.t.\ \ \lambda_1 > 0, \lambda_2 > 0 \end{aligned} \tag{6}$$

where the second term controls the sparsity in the feature level, and the third term controls the sparsity in the group level. If $\lambda_2 = 0$, the above problem degenerates to LASSO. When $\lambda_1 = 0$, it degenerated to the group LASSO.

3.2 Sparse technique for uncertain feature selection

Through the above subsection, we know that the LASSO does not consider the group information and selects a subset of features from all groups. The elastic net selects groups of correlated variables. The group LASSO selects a subset of the groups. The sparse group LASSO simultaneously selects a subset of the groups and a subset of the features within each selected group.

In uncertain data, x_{ij} is described by s values which form a PDF. Each sample x_i includes d features. Therefore, the features of uncertain data form a natural group structure. Each feature corresponds to a group and each group has s sub-features. When we want to select features for uncertain data, it is desirable to treat each feature with s sub-features as a unit when selecting important features. And we want to remove some redundant points in each feature. Based on this intension, sparse group LASSO should be chosen to solve this problem.

If we take the weights $w = \{w_1, w_2, \ldots, w_d\}$ as d features of uncertain data, each w_i includes s values. Here we set $w_i = \{w_{i1}, w_{i2}, \ldots, w_{is}\}$. The sparse group LASSO can be written as

$$\begin{aligned} &\min_w \tfrac{1}{2}||Xw - Y||_2^2 + \lambda_1||w||_1 + \lambda_2 \sum_{i=1}^{d} ||w_i||_2 \\ &s.t.\ \ \lambda_1 > 0, \lambda_2 > 0 \end{aligned} \qquad (7)$$

After the weights w of features and sub-features are obtained, the simplest method is that only the features and sub-features with nonzero weights are selected. Here we use $\widehat{w}$ to describe that the feature is selected or not. The size of $\widehat{w}$ is the same as that of $w.\widehat{w}$ can be obtained by

$$\widehat{w_{ij}} = \begin{cases} 0 & if\ w_{ij} = 0 \\ 1 & others \end{cases},$$

where $i = 1, 2, \ldots, d$ and $j = 1, 2, \ldots, s$. When the sum of absolute of w_i is equal to zero, the weights of the i-feature are all zero. Therefore, the i-feature should be removed. For other cases, some sub-features corresponds to zero weights should also be removed in each feature. After feature selection, the features X are changed into $X\widehat{w}$.

We define two indexes, number of features (FNum) and ratio as follows.

$$\text{FNum} = d - \sum_{i=1}^{d} \text{ceil}\left(\frac{\sum_{j=1}^{s} \widehat{w_{ij}}}{s}\right),$$

$$ratio = \frac{\sum_{i=1}^{d}\sum_{j=1}^{s} \widehat{w_{ij}}}{d * s}$$

where $ceil(x)$ rounds the elements of x to the nearest integers towards infinity. FNum shows the relevant number of features in uncertain data. That is, FNum represents the degree of sparsity between groups. The greater FNum is, the greater the relevant number is, and the smaller the sparsity degree between groups is. Ratio shows the degree of sparsity in the groups. The less ratio is, the greater the sparsity degree in the groups. For feature selection, we want to get good accuracy with less FNum and ratio.

For the optimal problem of sparse group LASSO in Eq. 7, we can see that the optimization problem is the sum of convex functions [27]. The first term is the squared loss which is smooth. The last two terms are the regularizer which is non-smooth [6]. The algorithm for this optimal problem has been explored in [15] and [21]. We employ SLEP software [19] to apply the feature selection for

Table 1 Data sets

Number	Data	Samples	Features	Classes
1	Crx(Credit)	690	15	2
2	German	1,000	20	2
3	Heart	270	13	2
4	Hepatitis	155	19	2
5	Horse	368	22	2
6	Iono	351	32	2
7	Sonar	208	60	2
8	WDBC	569	30	2
9	WPBC	198	33	2

Table 2 Accuracy with HIK SVM for uncertain data with Gaussian noise

Data	Raw data	Uncertaindata(w, s)	FS	FS(ratio < 0.9)
Crx	84.91 ± 18.02	85.93 ± 12.50 (0.01, 10)	86.07 ± 12.60 (0.01, 10)	84.62 ± 17.53 (0.20, 100)
German	74.70 ± 5.31	75.80 ± 4.49 (0.10, 90)	75.80 ± 4.49 (0.10, 90)	75.50 ± 4.77 (0.10, 100)
Heart	86.67 ± 6.34	86.30 ± 6.31 (0.10, 20)	86.30 ± 6.31(0.10, 20)	84.07 ± 6.99 (0.05, 80)
Hepatitis	90.50 ± 6.58	92.17 ± 5.33 (0.05, 20)	94.17 ± 5.84 (0.01, 50)	92.83 ± 5.88 (0.01, 60)
Horse	94.32 ± 3.87	93.79 ± 4.70 (0.20, 60)	95.37 ± 3.90 (0.01, 10)	95.37 ± 3.90 (0.01, 10)
Iono	92.69 ± 5.85	92.96 ± 6.00 (0.10, 10)	92.96 ± 6.00 (0.10, 10)	92.67 ± 6.82 (0.20, 30)
Sonar	82.64 ± 9.86	86.07 ± 8.69 (0.20, 10)	88.48 ± 7.62 (0.10, 10)	88.00 ± 7.30 (0.05, 20)
WDBC	97.89 ± 1.81	98.60 ± 1.38 (0.10, 30)	98.60 ± 1.38 (0.10, 30)	98.07 ± 1.29 (0.01, 30)
WPBC	83.42 ± 8.17	81.42 ± 8.74 (0.20, 70)	81.92 ± 10.83 (0.20, 50)	79.79 ± 10.00 (0.20, 100)
Ave.	87.43	88.12	88.85	87.88

uncertain data. The SLEP package is available online and has done an efficient implementation of the algorithm of paper [21] in a matlab interfaced module. This algorithm is a sub-gradient based approach. It iteratively computes the gradient update. At each iteration, the first-order black-box method [23] is used. Thus only the function value and gradient are needed to evaluate and the convergence rate is optimal for smooth convex optimization. The Euclidean projections can be computed either analytically or in linear time [20]. Therefore, this procedure can be applied to large datasets. These properties make the procedure suitable to select the features of uncertain data.

Here we briefly review the computation for the sparse group LASSO [11, 21]. The algorithm first focuses on one group i and then iterates this step over groups $i = 1, 2, \ldots, d$ until convergence. Let $X_i = Z = (Z_1, Z_2, \ldots, Z_s)$ denote the data in group i, the weights $w_i = \theta = (\theta_1, \theta_2, \ldots, \theta_s)$, and the residual $r_i = y - \sum_{k \neq i} X_k w_k$.

Table 3 Parameters and number of features for uncertain data with Gaussian noise

Data	FS of features				FS of features(ratio $<$ 0.9)			
	FNum	ratio	λ_1	λ_2	FNum	ratio	λ_1	λ_2
Crx	15.0	0.9960	100	0.1	8.9	0.5750	1,000	0
German	20.0	1.0000	10	10	16.1	0.6864	1,000	1,000
Heart	13.0	1.0000	10	10	5.3	0.3943	1,000	1,000
Hepatitis	17.9	0.9287	100	0	17.0	0.8947	0	1,000
Horse	16.6	0.7545	0	1,000	16.6	0.7545	0	1,000
Iono	32.0	1.0000	10	10	27.7	0.8756	0	1,000
Sonar	54.4	0.7172	100	0.1	51.0	0.7396	100	0.10
WDBC	30.0	1.0000	10	10	24.8	0.8267	0	1,000
WPBC	33.0	0.4291	100	0	18.3	0.5545	0	1,000
Ave.	25.8	0.8684			20.6	0.6957		

Table 4 Accuracy with HIK SVM for uncertain data with uniform noise

Data	Uncertain data(w, s)	FS	FS(ratio $<$ 0.9)
Crx	85.22 ± 14.27 (0.01, 60)	85.22 ± 14.27 (0.01, 60)	84.91 ± 17.17 (0.20, 90)
German	75.80 ± 3.68 (0.05, 90)	75.80 ± 3.68 (0.05, 90)	74.90 ± 4.63 (0.01, 50)
Heart	86.30 ± 6.54 (0.10, 30)	86.67 ± 6.81 (0.10, 30)	83.70 ± 5.00 (0.05, 70)
Hepatitis	92.17 ± 5.33 (0.05, 10)	93.50 ± 5.47 (0.01, 30)	92.83 ± 5.88 (0.05, 40)
Horse	94.04 ± 4.14 (0.10, 20)	95.11 ± 3.33 (0.01, 50)	95.11 ± 3.33 (0.01, 50)
Iono	92.69 ± 5.85 (0.10, 20)	92.96 ± 6.81 (0.20, 60)	92.65 ± 5.73 (0.05, 30)
Sonar	87.05 ± 8.50 (0.20, 50)	87.55 ± 7.86 (0.01, 30)	87.55 ± 7.86 (0.01, 30)
WDBC	98.25 ± 1.65 (0.10, 20)	98.42 ± 1.29 (0.01, 30)	98.25 ± 1.43 (0.01, 30)
WPBC	81.92 ± 8.47 (0.20, 100)	81.92 ± 8.47 (0.20, 100)	80.34 ± 5.91 (0.10, 50)
Ave.	88.15	88.57	87.81

Table 5 Parameters and number of features for uncertain data with uniform noise

Data	FS of features				FS of features(ratio $<$ 0.9)			
	FNum	ratio	λ_1	λ_2	FNum	ratio	λ_1	λ_2
Crx	15.0	1.0000	10	10	8.9	0.5787	1,000	0
German	20.0	1.0000	0.1	0.1	16.2	0.7651	1,000	0
Heart	13.0	0.9974	100	0	7.2	0.4284	1,000	0
Hepatitis	17.9	0.9346	100	0.01	16.7	0.8789	0	1,000
Horse	11.4	0.4650	1000	0	11.4	0.4650	1,000	0
Iono	30.6	0.9562	0	1000	26.6	0.8312	0	1,000
Sonar	47.1	0.7297	100	0.1	47.1	0.7297	100	0.10
WDBC	30.0	0.9771	100	0.1	24.3	0.8100	0	1,000
WPBC	33.0	0.9997	10	10	28.2	0.3946	100	0
Ave.	24.2	0.8955			20.7	0.6535		

According to [11], the subgradient equations of Eq. 7 with respect to θ_j are

$$-Z_j^T\left(\sum_j Z_j\theta_j - r\right) + \lambda_1 t_j + \lambda_2 s_j, \quad j = 1, 2, \ldots, s$$

where

$$s_j = \begin{cases} \frac{\theta_j}{||\theta||_2}, & if\ \theta_i \neq 0 \\ ||s||_2 < 1, otherwise \end{cases},$$

$$t_j \begin{cases} = sign(\theta_j), & if\ \theta_j \neq 0 \\ \in [-1, 1], & otherwise \end{cases}.$$

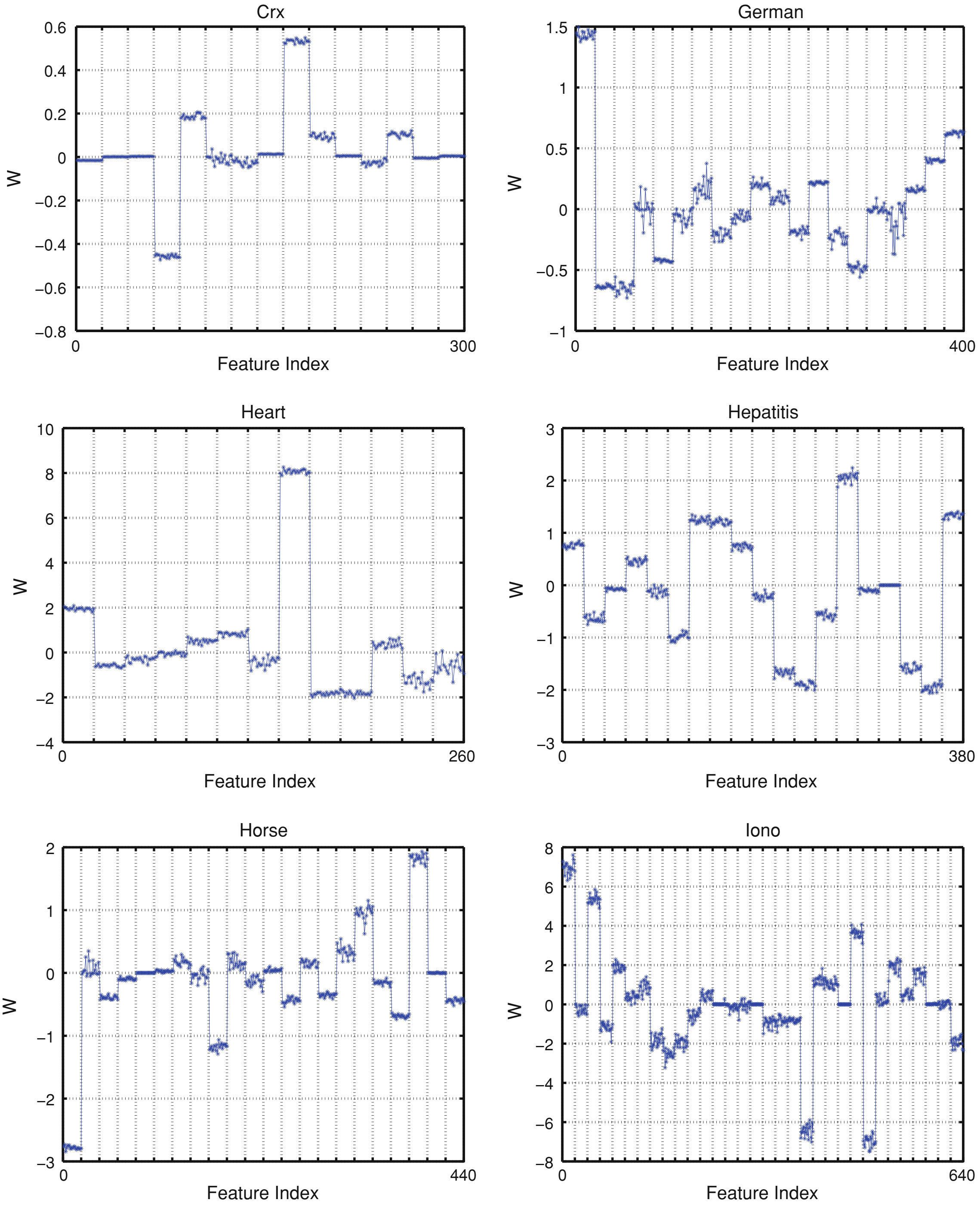

Fig. 2 Weights of nine datasets with Gaussian noise when $w = 0.05$, $s = 20$, $\lambda_1 = 100$, and $\lambda_2 = 0.1$ for the first six datasets

Let $\alpha = \boldsymbol{X_i r}.\boldsymbol{\theta} = \boldsymbol{0}$ only if $a_j = \alpha_1 t_j + \alpha_2 s_j$ has a solution with $\|s\|_2 < 1$ and $t_j \in [-1, 1]$. This can be determined by minimizing

$$J(t) = (1/\lambda_2{}^2)\sum_{j=1}^{s}(aj - \lambda_1 t_j)^2 = \sum_{j=1}^{s} s_j{}^2.$$

We denote $\widehat{*}$ as the solution of variant. If $J(\widehat{t}) \le 1, w_i = 0$. If $J(\widehat{t}) > 1$, we should minimize the criterion

$$\frac{1}{2}\sum_{i=1}^{l}\left(r_i - \sum_{j=1}^{s} Z_{ij}\theta_j\right)^2 + \lambda_1 \sum_{j=1}^{s}|\theta| + \lambda_2\|\theta\|_2.$$

The coordinate descent can be used to obtain the global minimum. Therefore, for $j = 1, 2..., s, \widehat{\theta_j} = 0$, if $|Z_j^T r_j| < \lambda_1$; otherwise minimizing

$$\frac{1}{2}\sum_{i=1}^{l}\left(y_i{}' - \sum_{j=1}^{s} Z_{ij}\theta_j\right)^2 + \lambda_1 \sum_{j=1}^{s}|\theta| + \lambda_2\|\theta\|_2$$

over θ_j by a one-dimensional optimization.

4 Experiments

To test the effectiveness of the proposed technique, experiments are carried out on nine datasets from the UCI machine learning repository [2]. The information about these datasets are summarized in Table 1.

HIK SVM is used as the classifier, which is realized by the software fast-additive-svms [22]. We conduct ten-fold cross-validation to compute the performance. The dataset is divided into ten subsets at first. Then nine subsets are used to train the model for feature selection. The features with nonzero weights are selected in the nine subsets to model HIK SVM. The left subset is used as the test dataset. We repeated the above process ten times until every subset has been used as the test sample. Finally, the average of the ten accuracies is recorded. Furthermore, we try various parameters in the experiment for w, s, and (λ_1,λ_2). Here, $w = [0.01, 0.05, 0.1, 0.2], s = [10, 20, 30, \ldots, 100]$,

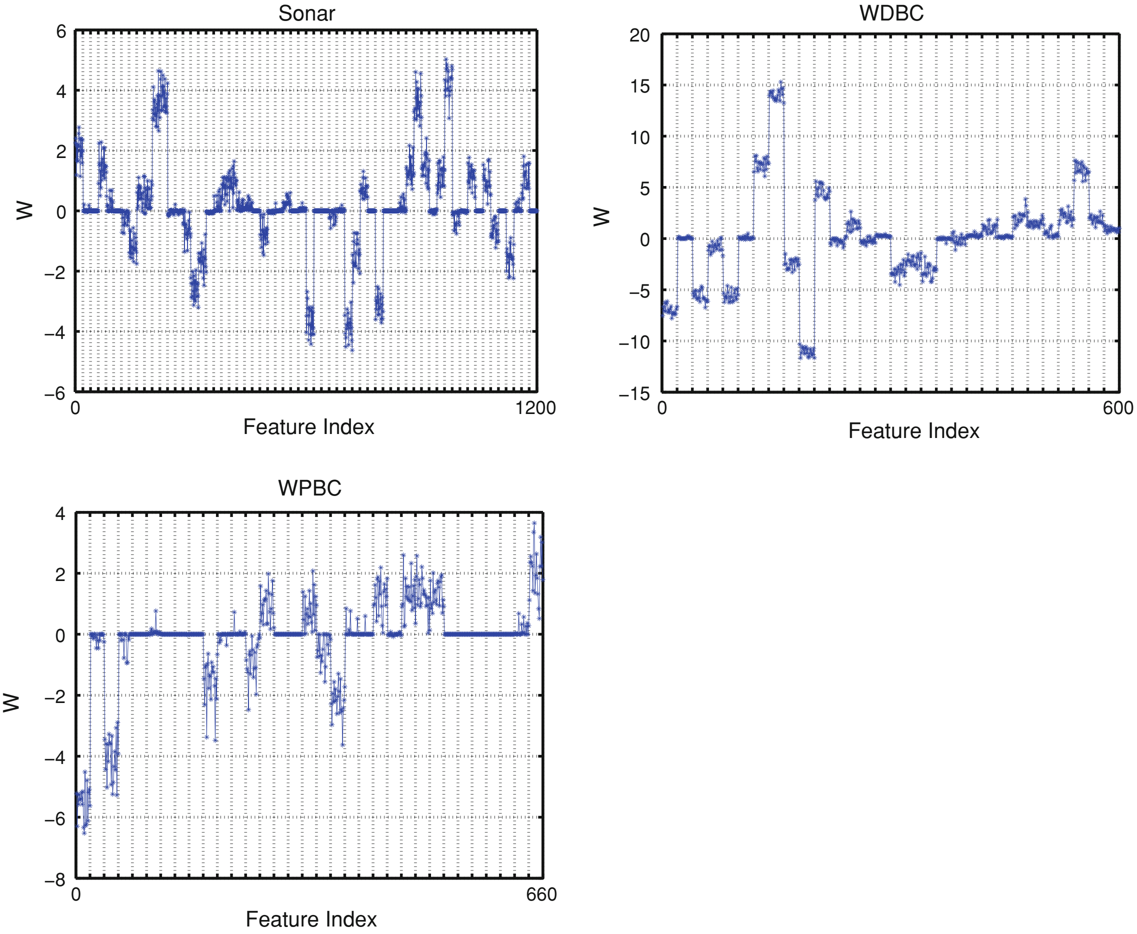

Fig. 3 Weights of nine datasets with Gaussian noise when $w = 0.05$, $s = 20$, $\lambda_1 = 100$, and $\lambda_2 = 0.1$ for left three datasets

and $(\lambda_1, \lambda_2) = \{(10^3, 10^3), (10\ , 10), (0.1, 0.1), (10^3, 0), (0, 10^3), (100, 0), (0, 100), (100, 0.1), (100, 10), (10, 100), (1, 0.001), (100, 0.01)\}$.

The results for Gaussian noise are given in Tables 2 and 3. We can see that the accuracy of uncertain data is greater then raw data. After feature selection, the best accuracy is improved furthermore. However, the feature number for the best accuracy is not changed much. Compared with Table 1, we can know that feature number for Crx, German, Heart, Iono, WDBC, WPBC is equal to the original data. In specially, the ratio of German, Heart, Iono and WDBC is 1 and the accuracy is also not changed. That is to say, feature selection is useless for these four datasets. For the other two datasets, Crx and WPBC, the ratio is 0.9960 and 0.4291, respectively. This shows that although the number of features is not change, some values in features are removed. In particularly, most values of features in WPBC are removed because the ratio is 0.4291, which is less than 0.5. Furthermore, the accuracy is improved with less features. This demonstrates that feature selection in the group is effective. For the left three datasets, Hepatitis, Horse, and Sonar, feature selection in and between groups is both effective. The number of features is less than the original dataset, the ratio is less than 1, and the average accuracy rises about 1.5 %.

For the best accuracy of feature selection, we can know that feature selection is effective for five datasets. Therefore, we select the results for the best accuracy of the ratio <0.9. We can see that the accuracy is a little lower than uncertain data, but the number and ratio of features are much less than the original one. The average feature number is 20.6, less than about 5 compared with that of the best accuracy 25.8. For the case of ratio <0.9, we can obtain comparative accuracy with 69.57 % of features. This shows that sparse group LASSO is effective both in and between groups of features. For parameters λ_1 and λ_2, FS (ratio $<$ 0.9) is greater than FS.

Tables 4 and 5 show the case of uncertain data with uniform noise. The derived conclusions are similar to the case of Gaussian noise.

Figures 2, 3, 4, 5 show the weights of features with different datasets and different parameters. Figures 2 and 3 show the weights of all nine datasets with $w = 0.05$, $s = 20$, $\lambda_1 = 100$, and $\lambda_2 = 0.1$. Nine subsets of each dataset are used to construct sparse group LASSO model. Each column in the figure shows the s sub-features of each

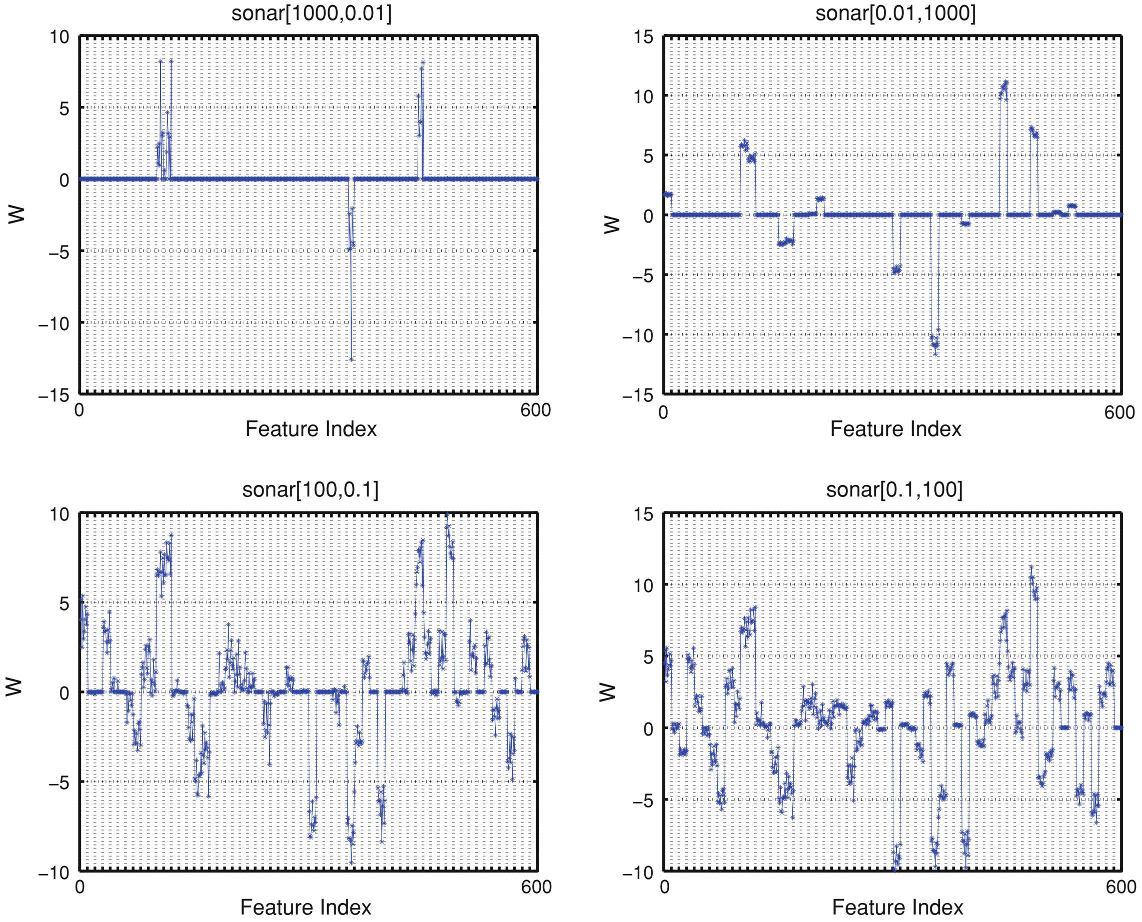

Fig. 4 Weights of Sonar with different λ_1, λ_2 and $w = 0.05$, $s = 10$. $[\lambda_1, \lambda_2]$ shows in the title for the first four subfigures

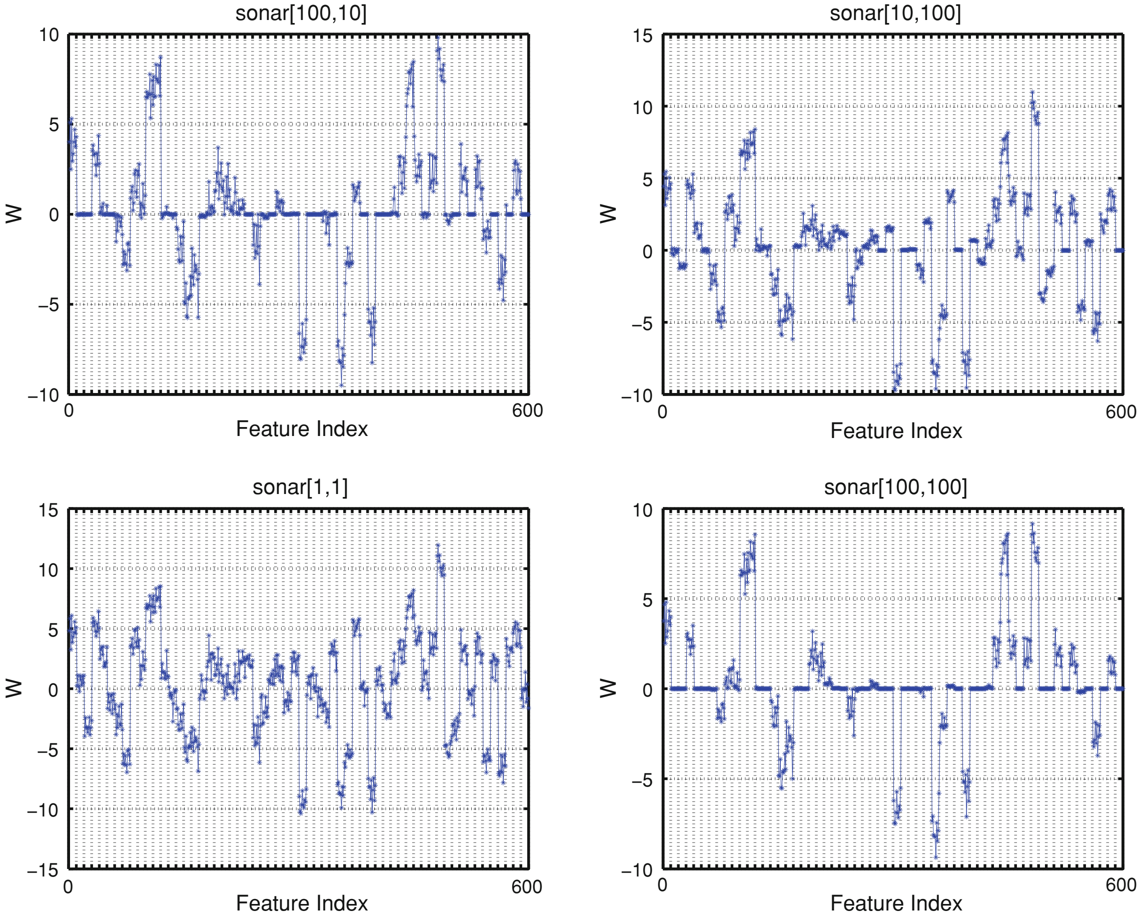

Fig. 5 Weights of Sonar with different λ_1, λ_2 and $w = 0.05$, $s = 10$. [λ_1, λ_2] shows in the title for the left four subfigures

feature. We can see that all the weights of some groups are zero. These features can be removed as a whole. Some weights in the group are also zero. However, because $\lambda_2 = 0.1$ is little, only a little of weights in the group are zero. Figures 4 and 5 show the weights of Sonar with different λ_1,λ_2 and $w = 0.05$,$s = 10$. Compared with the first two subfigures in Fig. 4, we can get that when λ_1 is great and λ_2 is small, the weights are very sparse in the groups; when λ_1 is small and λ_2 is great, the weights in the group are not sparse but that between groups are sparse. With the decrease of maximum of λ_1 and λ_2, the number of zero weights decreases. When $\lambda_1 = 100$ it is clear that the number of groups with all zero weights is much. The parameters λ_1 and λ_2 in Tables 3 and 5 show that if we want to get good accuracy with less feature the value of λ_1 and λ_2 should be large.

5 Conclusions and future work

In this paper, sparse group LASSO has been introduced for constructing a feature selection algorithm for uncertain data. Each feature in uncertain data is represented by s points, which form a PDF. Therefore, the features of uncertain data naturally have the structure of grouping. Each group includes s sub-features. The feature selection with sparse group LASSO can simultaneously encourage sparsity at the level of both features and sub-features. Experimental results on nine UCI databases containing Gaussian and uniform uncertainty show that the sparse group LASSO is effective to select relevant features and sub-features. We can get competent performance with about 70 % features as all features.

Future work could be focused on the two directions. First, we just discuss the binary classification with uncertain data. Although the algorithm can be easily extended to the case of multiple classes, the performance of the algorithm in multi-class tasks should be systematically discussed. Second, features are selected with a linear model in this work. We will generalized it to nonlinear tasks in the future.

Acknowledgments This work is supported by the National Natural Science Foundation of China under Grants 61105054, 61071179 and 10978011. And this article was partly supported by Key Laboratory of Network Oriented Intelligent Computation, Program for New Century Excellent Talents in University (No. NCET-08-0156); Dr Xie is supported by China Postdoctoral Science Foundation funded project.

References

1. Aggarwal C, Yu P (2009) A survey of uncertain data algorithms and applications. IEEE Trans Knowl Data Eng 21(5):609–623
2. Asuncion A, Newman D (2007) Uci machine learning repository [http://www.ics.uci.edu/∼mlearn/mlrepository.html]. Irvine, CA: University of california. School of Information and Computer Science
3. Bernecker T, Kriegel H, Renz M, Verhein F, Zuefle A (2009) Probabilistic frequent itemset mining in uncertain databases. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp. 119–128
4. Bernecker T, Kriegel H, Renz M, Verhein F, Züfle A (2012) Probabilistic frequent pattern growth for itemset mining in uncertain databases. In: Scientific and Statistical Database Management. Springer, Berlin, pp. 38–55
5. Bi J, Zhang T (2004) Support vector classification with input data uncertainty. Adv Neural Info Process Syst 17(5):161–168
6. Chatterjee S, Steinhaeuser K, Banerjee A, Chatterjee S, Ganguly A (2012) Sparse group lasso: consistency and climate applications. SDM
7. Cheng R, Chau M, Garofalakis M, Yu J (2010) Guest editors' introduction: special section on mining large uncertain and probabilistic databases. IEEE Trans Knowl Data Eng 22(9): 1201–1202
8. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, Cambridge
9. Doquire G, Verleysen M (2011) Feature selection with mutual information for uncertain data. Data Warehous Knowl Discov pp 330–341
10. Fletcher A, Rangan S, Goyal V (2009) Necessary and sufficient conditions for sparsity pattern recovery. IEEE Trans Info Theory 55(12):5758–5772
11. Friedman J, Hastie T, Tibshirani R (2010) A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736
12. Guyon I., Elisseeff A. (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
13. Hu Q, Pan W, An S, Ma P, Wei J (2010) An efficient gene selection technique for cancer recognition based on neighborhood mutual information. Int J Mach Learn Cybern 1(1):63–74
14. Huang J, Zhang T (2010) The benefit of group sparsity. Ann Stat 38(4):1978–2004
15. Jenatton R, Mairal J, Obozinski G, Bach F (2010) Proximal methods for sparse hierarchical dictionary learning. In: Proceedings of the international conference on machine learning (ICML)
16. Kanagal B, Deshpande A (2008) Online filtering, smoothing and probabilistic modeling of streaming data. In: IEEE 24th international conference on data engineering (ICDE) pp 1160–1169
17. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society conference on computer vision and pattern recognition, vol 2, pp 2169–2178
18. Lian X, Chen L (2012) Probabilistic top-k dominating queries in uncertain databases. Inf Sci
19. Liu J, Ji S, Ye J Slep (2009) Sparse learning with efficient projections. Arizona State University, Glendale
20. Liu J, Ye J (2009) Efficient euclidean projections in linear time. In: Proceedings of the 26th annual international conference on machine learning. ACM, New York, pp 657–664
21. Liu J, Ye J (2010) Moreau-yosida regularization for grouped tree structure learning. Adv Neural Info Process Syst 23:1459–1467
22. Maji S, Berg A, Malik J (2008) Classification using intersection kernel support vector machines is efficient. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 1–8
23. Nesterov Y (2003) Introductory lectures on convex optimization: a basic course, vol 87. Springer, Berlin
24. Peng J, Zhu J, Bergamaschi A, Han W, Noh D, Pollack J, Wang P (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. Ann Appl Stat 4(1):53–77
25. Qin X, Zhang Y, Li X, Wang Y (2010) Associative classifier for uncertain data. In: Web-Age Information Management, pp 692–703
26. Ren J, Lee S, Chen X, Kao B, Cheng R, Cheung D (2009) Naive bayes classification of uncertain data. In: Ninth IEEE international conference on data mining. IEEE Computer Society, Washington, pp. 944–949
27. Rockafellar R (1996) Convex analysis, vol. 28. Princeton university press, Princeton
28. Sharma A., Imoto S., Miyano S., Sharma V. (2011) Null space based feature selection method for gene expression data. Int J Mach Learn Cybern pp 1–8
29. Shivaswamy P, Bhattacharyya C, Smola A (2006) Second order cone programming approaches for handling missing and uncertain data. J Mach Learn Res 7:1283–1314
30. Subrahmanya N, Shin Y (2012) A variational bayesian framework for group feature selection. Int J Mach Learn Cybern pp 1–11
31. Tang V., Yan H. (2012) Noise reduction in microarray gene expression data based on spectral analysis. Int J Mach Learn Cybern 3(1):51–57
32. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodological), pp 267–288
33. Tong Y, Chen L, Cheng Y, Yu P (2012) Mining frequent itemsets over uncertain databases. Proc VLDB Endow 5(11):1650–1661
34. Tsang S, Kao B, Yip K, Ho W, Lee S (2011) Decision trees for uncertain data. IEEE Trans Knowl Data Eng 23(1):64–78
35. Yuan M, Lin Y (2005) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Statistical Methodology) 68(1):49–67
36. Zou H., Hastie T. (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodology) 67(2):301–320