

A Novel Automatic Grouping Algorithm for Feature Selection

Qiulong Yuan and Yuchun Fang^(✉)

School of Computer Engineering and Science, Shanghai University,
Shanghai, China
ycfang@shu.edu.cn

Abstract. Feature selection is used in many application areas relevant to expert and intelligent system such machine learning, bioinformatics and image processing. Feature selection plays an important role in reducing the dimensionality of high-dimensional features. However, traditional feature selection methods are not able to intelligently learn intrinsic data structures. In this paper, we proposed a novel feature selection method, which can automatically learn grouping structure relation among features. Experiments are conducted on the selection of both raw features and statistically handled features. Experimental results demonstrate that the proposed method can identify important features by automatic grouping, and outperforms the other methods on several public data sets. Moreover, by using parallel computing, the training time consumed by our method is only 50% of that of the traditional methods.

Keywords: Feature selection · Mutual information
Automatic grouping

1 Introduction

Feature selection is used in many application areas relevant to expert and intelligent system such machine learning, bioinformatics and image processing. In computer vision, machine learning, and data mining, data are always represented by high-dimensional feature vectors. But high-dimensional data can increase the consumption of processing time and storage space during processing. Moreover, most of the existing machine learning methods, such as classification, regression, or other tasks, are mainly designed more adaptive to low-dimensional data. The computation of high-dimensional data can usually become much more complex and difficult. As a solution to this issue, feature selection (also known as variable selection) [2, 13, 20, 21] is performed to choose a representative subset from the high dimensional features. The subset is expected to bear sufficient information about the original high-dimensional feature set for specific learning tasks. The purpose of feature selection is to find the relevant feature subset that best reflects the statistical properties of the pattern category to represent the original feature.

According to different evaluation metrics, feature selection algorithms can roughly be classified into three categories, i.e., filter, wrapper and embedded methods [13]. The filter methods rely on general characteristics of the data to evaluate and select feature subsets without considering the learning algorithm. Such as variance and Fisher score [11, 14]. The wrapper model takes the performance of the learner to be used as the evaluation criterion. For embedded methods, the process of feature selection and learner training is integrated, both are completed in the same process, and in the process of training the learning algorithms automatic feature selection.

Recently, some new embedded feature selection methods that integrate the theory of sparse representation [3, 8, 24], compressed sensing, and feature selection [4, 18, 26] have been proposed. Sparsity-inducing feature selection methods have been widely used in face authentication [5, 23], face detection, face attributes classification [9], and gene expression [25]. However, L_1 regularization computation typically requires solving either NP -problem or an alternative problem that still involves a costly iterative optimization [7].

In practical applications, the features have some essential structures. Integrating knowledge about the feature structures may help identify the important features. Ye and Liu [27] and Zhang et al. [28] proposed to use the group structure information of the data to carry on the feature selection. But the existence of these methods have a defect that the number of groups in the group structure is man-made, and the automatic grouping is not realized. On the one hand, the feature set F is artificially divided into k groups, and this grouping is usually carried out by experience, still do not achieve automatic grouping, on the other hand, the so-called grouping is to divide the adjacent m features into a group, the features that are adjacent to each other do not necessarily belong to the same group.

Taking into account these factors, in this paper, we propose a novel automatic grouping feature selection method, which use l_2 norm to ensure group effect [31] and use the mutual information measure to ensure the low redundancy within the group. The Laplacian regularization based on mutual information is used in the objective function of the method, so we call it MIL. Our experiments demonstrate the efficiency of the proposed method. The difference between our approach and the traditional group lasso method is shown in Fig. 1. As shown in Fig. 1, the traditional group lasso method is continuous and the number of groups is decided by people. In addition, the group is fixed once it is determined. The MIL method combines autonomously by judging the correlation between features, where each group of features can be discontinuous and the number of features of each group can be unequal.

The remainder of the paper is organized as follows. Sect. 2 presents different related methods and information theory used in our paper. Section 3 introduces our proposed method. Experimental evaluation is depicted in Sect. 4. At last, we conclude the paper in Sect. 5.

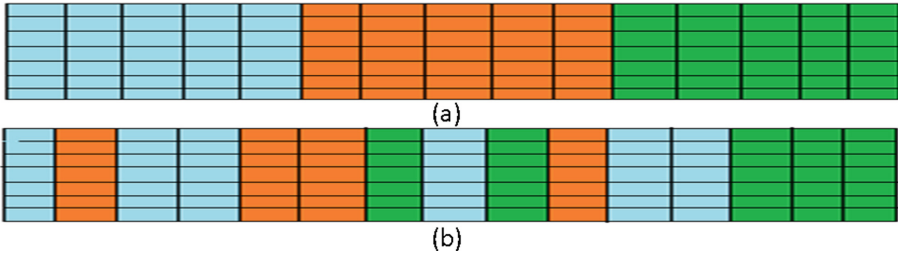


Fig. 1. Comparison between traditional group lasso and MIL methods. Suppose there is a data set $A \in R^{7 \times 15}$, we use the 7×15 table to represent the A dataset, each row represents a sample, each column represents a feature, the same color indicates the same group, (a) traditional group lasso grouping result, (b) MIL grouping result. (Color figure online)

2 Related Work

In this section, we will introduce some of the existing feature selection algorithms, and then introduce the basics of the information theory, as we will use it in Sect. 3. Before the details are presented, we need to summarize the notations and the definitions used in this paper. Let $A = [a_1, a_2, \dots, a_n] \in R^{n \times d}$ be n samples data in the d -dimensional space, where n is the number of sample and d is the number of features. Accordingly, denote label vector $y = [y_1, y_2, \dots, y_n]^T \in R^{n \times 1}$, which $y_i \in \{+1, -1\}, i \in \{1, 2, \dots, n\}$ if task is binary classification. Denote λ as the hyper parameter to balance the data misfit and the penalty. Denote $W = [w_1, w_2, \dots, w_d]^T \in R^{d \times 1}$ as the unknown weight coefficient vector, which need we to be estimated.

2.1 Related Feature Selection Methods

As a kind of embedded method, regularization techniques based on L_1 norm has been widely used to cope with feature selection in machine learning tasks. According to compressive sensing theory, the minimum L_1 norm solution to an under determined system of linear equation is equivalent to the sparsest possible solution under general conditions. Destrero et al. [5,6] used lasso for feature selection in face detection and face authentication. Its objective function is:

$$\min_W \|y - WA\|_2^2 + \lambda \|W\|_1 \tag{1}$$

Lasso is suboptimal since it produces biased estimates for the large coefficients. Zou [30] found that Lasso uses the same degree of compression for all coefficients. And Lasso does not have the oracle properties. In order to improve the performance of Lasso, the adaptive Lasso [30] is proposed

$$\min_W \|y - WA\|_2^2 + \lambda \sum_{i=1}^d a_i \|w_i\|_1 \tag{2}$$

From Eq. 2, we know the only difference between Lasso and adaptive Lasso is the latter gives a weight coefficient for each. Different compression coefficients are used for different weights, let Lasso has oracle properties. It not only has good usability in practice but also has an excellent character in theory. Obviously, when all coefficient is equal, the adaptive Lasso is equivalent to Lasso.

The fused Lasso introduced in [22] get a solution that has sparse both the coefficient and their successive differences. The objective function of fused Lasso can be represented as follows:

$$\min_W \|y - WA\|_2^2 + \|W\|_1 + \alpha \sum_{i=2}^d |w_i - w_{i+1}| \tag{3}$$

The bridge estimator [17] is defined as follows:

$$\min_W \|y - WA\|_2^2 + \lambda \sum_{i=1}^d |w_i|^\gamma \tag{4}$$

The bridge estimator has two important special cases. When $\gamma = 2$, it is popular ridge estimator. When $\gamma = 1$, it is the Lasso.

In many practical applications, some features often have a strong correlation. In this case, the lasso tends to select only one of the correlated features. To deal with feature with strong correlation, Zou and Hastie [31] proposed elastic net regularization as

$$\min_W \|y - WA\|_2^2 + \alpha \sum_{i=1}^d |w_i| + (1 - \alpha) \sum_{i=1}^d |w_i|_2 \tag{5}$$

where $|\cdot|_2$ is the L_2 -norm. [31] show that L_2 regularization has the group effect and L_1 regularization does not have group effect Zou et al. [31] add a group effect to lasso by using L_2 regularization to handle feature with strong correlations. Furthermore, when α is equal to 1, the elastic net is lasso, and when α is equal to 0, the elastic net is ridge estimator [17].

The penalties introduced in Eqs. (1)–(5) are assume that features are independent and ignored the structures of features completely [27]. However, in practical application, the features have some essential structures, such as groups [27, 28]. Suppose that features are divided into k groups. With the group structure, the W is rewritten as k groups $W = \{w_{G1}, w_{G2}, \dots, w_{Gk}\}$, and the objective function of group lasso as follows:

$$\min_W \|y - WA\|_2^2 + \sum_{i=1}^k \beta_i \|w_{Gi}\|_q \tag{6}$$

where the $\|\cdot\|_q$ is indicate q -norm, and β_i is the weight coefficient of i -th group. There are different structured feature selection methods according to different q values or different constraints, such sparse group lasso [29]. The group structure

provides good access to the structural property of the data. However, the number of groups in the group structure is man-made, and the automatic grouping is not realized. Moreover, the features that are adjacent to each other do not necessarily belong to the same group.

2.2 Information Theory

Given the two variables U, V , if their respective marginal probability distribution and joint probability distribution are respectively $p(u), p(v)$ and $p(u, v)$, and then their mutual information $I(u, v)$ is defined as:

$$I(u, v) = \sum_{u,v} p(u, v) \log \frac{p(u, v)}{p(u)p(v)} \tag{7}$$

When the variables U and V completely unrelated or independent of each other, the minimum mutual information, the result is zero, which means that there is no overlapping information between the two variables; on the other hand, the greater the interdependence, mutual information value will be greater.

3 MIL

In Sect. 2, we analyze some feature selection algorithms. On the one hand, L_1 regularization computation typically requires solving either NP -problem or an alternative problem that still involves a costly iterative optimization [7], on the other hand, Traditional methods based on structural features are not automatically group. Therefore, we proposed a novel feature selection model, MIL.

Assume given a set of training samples $A \in R^{n \times d}$ and the target labels $y \in R^{n \times 1}$ of the corresponding samples, the MIL uses the following criterion:

$$\min_W \|y - WA\|_2^2 + \alpha \|W\|_2 + \beta \sum_i \sum_j MI_{ij} (w_i - w_j)^2. \tag{8}$$

to find W , where α and β is the hyper parameter of the objective function. The first term of function (8) is the least-squares function. According to the least-squares method to seek the relationship of features and target labels. The second term of function (8) is l_2 norm. In this paper, the group effect of MIL is obtained via the L_2 norm, where the group is determined by the correlation between the features, and if the correlation between the two features is large, then we classify it as a group, and if the correlation is small, it is considered not a group. The third term of (8) is a manifold regularization where MI_{ij} is the correlation between the i -th and j -th feature. In this paper, the MI_{ij} is defined as the mutual information of the i -th and j -th feature (Mutual information can be used to measure the degree of interdependence between the two variables, and not limited to linear correlation, it also can be applied to nonlinear correlation). It is reasonable to require w_i and w_j close to each other if the i -th and j -th

feature Similarity (redundancy) is low, which is the objective of the term of the third term of (8). In MIL, we use the third term of (8) to ensure that the redundancy in the group is minimal. In fact, the third term of (8) have the ability of l_1 penalty in feature selection, it can ensure the minimal redundancy of the feature subset. Denote MI as the similarity matrix constructed by all MI_{ij} and the diagonal matrix D where the element i -th of is the sum of the i -th row of the MI . Therefore, we can get the Laplacian matrix $L = D - MI$ [15], and the third term of (8) can be represented as $\beta W^T L W$ by simple algebra. The objective function can be rewritten as follows:

$$\min_W \|y - WA\|_2^2 + \alpha \|W\|_2 + \beta W^T L W. \quad (9)$$

Fortunately, (9) has an analytical solution as follows:

$$W = (A^T A + \alpha I + \beta L)^{-1} A^T y. \quad (10)$$

After obtain W , we can rank features according to $|w_i|$. The larger $|w_i|$ is, the more important this feature is [12]. We can either select a fixed number of the most important features or set a threshold and select the feature whose $|w_i|$ is larger than the value [16].

4 Experimental Evaluation

In the section, we present experimental results of our method. The performance of the newly proposed method in this paper, MIL, is mainly compared with five other methods: Lasso (L_1 regularization) [5], Ridge (L_2 regularization) [17], elastic-net [31], and group lasso [28]. These methods are chosen for the following reasons: (a) Like MIL approach, the choice of these methods based on regularization; (b) These methods are reported in the paper to provide good performance; (c) These methods contain structure and non-structure method, group effect and no group effect method, it helps to compare the performance of the algorithms. In the experiment, we use non-image and image data set.

The non-image data sets are Colon [1] and Leukemia [10], where Colon contains 2000 dimensions raw features and Leukemia contains 7029 dimensions raw features. For image data set, we select the CFW [19] face data set, which is a large collection of celebrity face images from the Internet. The data set contains 200,000 face images for 1,500 celebrities. We selected 8,000 face images (20 images \times 400 people) from the CFW 60K to carry out attributes classification experiments, the 14 kinds of face attributes included in the selected pictures were gender, race, age and so on (more detail see Table 1). Then we use the ULBP to extracting low-level features. To be more specific, we scaling the face images to 140×160 pixels, and divided image into 7×8 cells, each of cell is 20×20 pixels. Then use the ULBP descriptor to extract 3,304-dimensional features as the raw features.

4.1 Performance Analysis with Colon and Leukemia Dataset

In the first experiment, we use the non-image data sets to evaluate our approach. We only selected the 50 features to observe the differences between the different methods. Figures 2 and 3 show the classification accuracy of the two datasets (Colon, Leukemia). As shown in Fig. 2, which illustrates the experiment with Colon, MIL reached stability with just 12 features and from the 4-th feature, MIL has always been far ahead of other methods. Compared to MIL, several other methods do not achieve better classification results. Without losing generality, the Fig. 3 shows similar results in Leukemia data set.

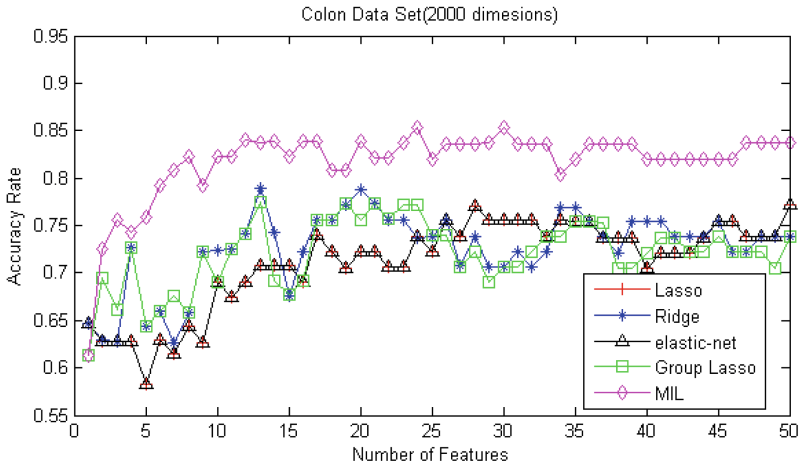


Fig. 2. Classification accuracy rate achieved with the Colon data set.

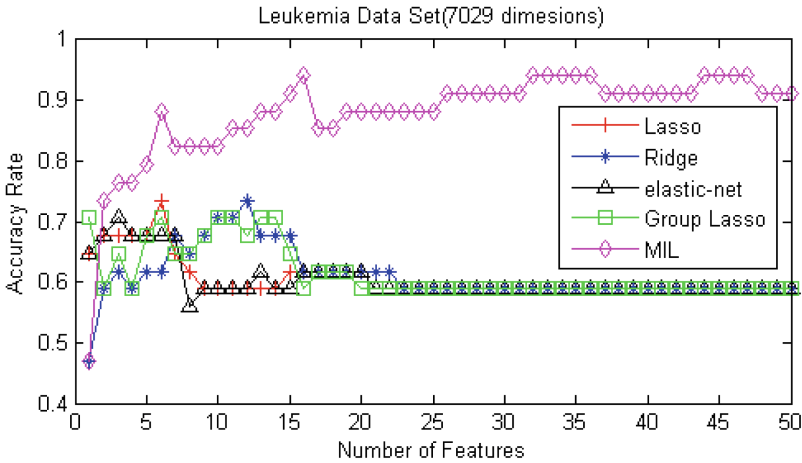


Fig. 3. Classification accuracy rate achieved with the Leukemia data set.

4.2 Performance Analysis on Face Dataset

In this experiment, we test MIL on image data set. In CFW dataset, we chose 14 face attributes (more detail see Table 1), and we used the MIL method to classify the 14 attributes. To observe the variation of accuracy with respect to a number of dimensions, the recognition rate is calculated from 10, 20, ..., to 100, and from 100, 200, ..., to 700. The results are shown in Table 1.

From the Table 1, we can find that our method has good performance. By calculating the average recognition rate of all dimensions, it can reflect the relationship between the recognition rate and the dimension. In general, if the average recognition rate is higher, then the number of dimensions required to achieve the same recognition rate will be less. However, the classification results obtained by the group sparse are not bad. We all know that most of the image feature description is based on statistical method, ULBP is the case, which itself counts as an area of information. In this article, the statistical size of the region is 20×20 pixels. Therefore, selecting continuous features as a set does not have much impact on the outcome of the experiment. The flaw in algorithms that cannot be grouped automatically is not so obvious.

Table 1. The average recognition accuracy rate (%) of five different methods, highest values are in bold, the average is calculated by the mean accuracy rate across a range of feature set size (from 10 to the maximum number (700) of selected features).

Attributes	Lasso	Ridge	Elastic-net	Group lasso	MIL
Man	79.52	89.11	89.57	89.40	90.42
Female	79.59	89.10	89.57	89.40	90.31
Asian	93.58	94.91	94.73	94.94	95.16
White	81.27	84.82	81.64	84.92	83.18
Black	90.88	93.54	94.59	93.60	93.12
Indian	93.52	95.12	94.59	95.14	95.24
Youth	66.28	71.71	64.33	71.87	69.88
Middlle age	67.12	69.11	67.25	69.25	66.88
Senior	90.79	93.20	92.37	93.26	93.34
No glasses	96.41	97.16	95.38	97.21	97.60
Eye glasses	97.43	97.50	96.66	97.54	98.01
Sun glasses	98.78	99.19	98.74	99.20	99.20
Positive expression	56.08	71.82	71.87	71.92	68.00
Neutral expression	55.01	71.64	71.87	71.92	68.14

4.3 Efficiency

Through the above analysis, we can say that our method is effective in feature selection. In this experiment, we test the efficiency of MIL compared to other

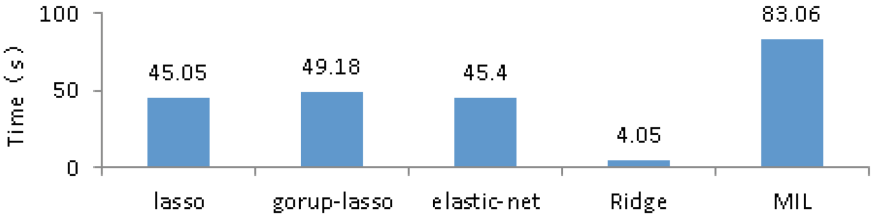


Fig. 4. Run time (s) comparison between MIL and other feature selection algorithms, measured by the time for computing W weight coefficient vector from gender attribute of the CFW data set.

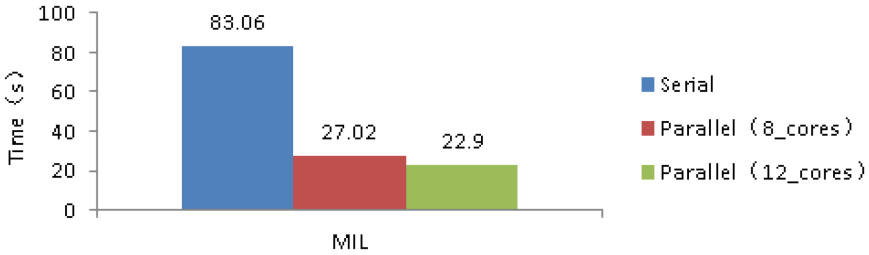


Fig. 5. Run time (s) comparison between serial and parallel MIL, measured by the time for computing W weight coefficient vector from gender attribute of the CFW data set.

approaches. However, we know the time complexity of MIL is $O(kNN)$, where N is defined as the feature numbers. MIL is thus generally more computationally demanding. In this paper, the raw features of face data set are 3304 dimensions. We use the time we calculate the weight coefficient W as a measure. In the experiment, which involves calculating L_1 regularization, we set the error tolerance to 10^{-6} . Taking face gender attribute as an example, the result is shown in Fig. 4.

As shown in Fig. 4, the proposed method takes 83.06s to calculate CFW dataset with 3304 dimensions features. MIL is considerably expensive. Nevertheless, we believe that the time complexity should not be a major deterrent to the practicality of MIL. There are many applications where the data collecting time is far more than the time required for data mining tasks such as feature selection (e.g., days to months for data collection vs. hours for data mining). Commodity multi-core systems are common nowadays, and it is straightforward to parallelize MIL to harness this parallel processing power. In these cases, it is justifiable to spend significant amounts of time for data processing and the improved performance brought about by MIL will be worth the effort. Towards this end, we tested a parallel version of MIL. The effectiveness of parallelization can be clearly observed in Fig. 5.

As seen in Fig. 5, the parallel version of MIL can effectively reduce the computation time of W weights. When using a 8 core processor, the computational efficiency of the MIL is greatly improved, with a computation time of 27.02s.

However, as the number of cores increases, the latter's promotion is not very obvious. After using parallel computing, the amount of time that MIL spends on computing the same dimensions training sample will be shortened to half of elastic-net and group-lasso.

5 Conclusions

In this paper, we proposed a novel feature selection method, which can realize automatic grouping. The proposed method is to use mutual information to achieve the minimum redundancy of each group. Because the amount of calculation of mutual information is increased by the increase in the number of features, we use parallel computing to reduce the cost of computing. It is worth mentioning that the calculation of L_1 regularization can not use parallel computing to speed up. We compare with other feature selection methods to evaluate the effectiveness of the proposed MIL. Experimental results show that the MIL can obtain high recognition rate with fewer feature dimensions, and outperforms the other methods on several public data sets.

Acknowledgement. The work is funded by the National Natural Science Foundation of China (Nos. 61371149, 61170155), Shanghai Innovation Action Plan Project (No. 16511 101200) and the Open Project Program of the National Laboratory of Pattern Recognition (No. 201600017)

References

1. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci.* **96**(12), 6745–6750 (1999)
2. Anne-Claire, H., Pierre, G., Jean-Philippe, V.: The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One* **6**(12), e28210 (2011)
3. Cheng, H., Liu, Z., Yang, L., Chen, X.: Sparse representation and learning in visual recognition: theory and applications. *Sig. Process.* **93**(6), 1408–1425 (2013)
4. Cong, Y., Wang, S., Liu, J., Cao, J., Yang, Y., Luo, J.: Deep sparse feature selection for computer aided endoscopy diagnosis. *Pattern Recogn.* **48**(3), 907–917 (2015)
5. Destro, A., De Mol, C., Odone, F., Verri, A.: A sparsity-enforcing method for learning face features. *IEEE Trans. Image Process.* **18**(1), 188 (2009)
6. Destro, A., De Mol, C., Odone, F., Verri, A.: A regularized approach to feature selection for face detection. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) *ACCV 2007. LNCS*, vol. 4844, pp. 881–890. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76390-1_86
7. Donoho, D.L.: Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
8. Elad, M.: *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, New York (2010). <https://doi.org/10.1007/978-1-4419-7011-4>

9. Fang, Y., Chang, L.: Multi-instance feature learning based on sparse representation for facial expression recognition. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015. LNCS, vol. 8935, pp. 224–233. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-14445-0_20
10. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**(5439), 531–537 (1999)
11. Gu, Q., Li, Z., Han, J.: Generalized fisher score for feature selection, pp. 266–273 (2012)
12. Gui, J., Sun, Z., Ji, S., Tao, D., Tan, T.: Feature selection based on structured sparsity: a comprehensive study. *IEEE Trans. Neural Netw. Learn. Syst.* **PP**(99), 1–18 (2016)
13. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**(6), 1157–1182 (2003)
14. Guyon, I.: Pattern classification. *Pattern Anal. Appl.* **1**(2), 142–143 (1998)
15. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face recognition using Laplacian-faces. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(3), 328–340 (2005)
16. Hou, C., Nie, F., Yi, D., Wu, Y.: Feature selection via joint embedding learning and sparse regression. In: Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI 2011, Barcelona, Catalonia, Spain, July, pp. 1324–1329D (2011)
17. Huang, J., Horowitz, J.L., Ma, S.: Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Stat.* **36**(2), 587–613 (2008)
18. Kim, Y., Kim, J.: Gradient lasso for feature selection. In: International Conference on Machine Learning, p. 60 (2004)
19. Li, Y., Wang, R., Liu, H., Jiang, H.: Two birds, one stone: jointly learning binary code for large-scale face image retrieval and attributes prediction. In: IEEE International Conference on Computer Vision, pp. 3819–3827 (2015)
20. Saeyns, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**(19), 2507 (2007)
21. Tang, J., Alelyani, S., Liu, H.: Feature selection for classification: a review. In: Documentacin Administrativa, pp. 313–334 (2014)
22. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **67**(1), 91–108 (2005)
23. Vo, N., Moran, B., Challa, S.: Nonnegative-least-square classifier for face recognition. In: Yu, W., He, H., Zhang, N. (eds.) ISSN 2009. LNCS, vol. 5553, pp. 449–456. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01513-7_49
24. Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T.S., Yan, S.: Sparse representation for computer vision and pattern recognition. *Proc. IEEE* **98**(6), 1031–1044 (2010)
25. Hang, X., Wu, F.X.: Sparse representation for classification of tumors using gene expression data. *J. Biomed. Biotechnol.* **2009**(1), 403689 (2009)
26. Yan, H., Yang, J.: Sparse discriminative feature selection. *Pattern Recogn.* **48**(5), 1827–1835 (2015)
27. Ye, J., Liu, J.: Sparse methods for biomedical data. *ACM* (2012)
28. Zhang, S., Huang, J., Li, H., Metaxas, D.N.: Automatic image annotation and retrieval using group sparsity. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **42**(3), 838–849 (2012). A Publication of the IEEE Systems Man & Cybernetics Society

29. Zhou, J., Liu, J., Narayan, V.A., Ye, J.: Modeling disease progression via fused sparse group lasso. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1095–1103 (2012)
30. Zou, H.: The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**(476), 1418–1429 (2006)
31. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. R. Stat. Soc.* **67**(2), 768–768 (2005)