Pattern Recognition 81 (2018) 388–403

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Feature weight estimation based on dynamic representation and neighbor sparse reconstruction

Xiaojuan Huang, Li Zhang*, Bangjun Wang, Zhao Zhang, Fanzhang Li

School of Computer Science and Technology & Joint International Research Laboratory of Machine Learning and Neuromorphic Computing, Soochow University, Suzhou, Jiangsu 215006, China

ARTICLE INFO

Article history:
Received 29 June 2017
Revised 8 January 2018
Accepted 20 March 2018
Available online 13 April 2018

Keywords:
Feature weighting
Feature selection
Relief
Sparse learning
Local hyperplane
l_1 regularization
Classification

ABSTRACT

Relief-like algorithms have been widely used as feature selection to reduce the dimension of high-dimensional data which involves thousands of irrelevant variables because of their low computational cost and high accuracy. Classical Relief algorithms have not exactly shown the dynamic procedure that updates weight iteratively. This paper proposes an innovative feature weight estimation method, called dynamic representation and neighbor sparse reconstruction-based Relief (DRNSR-Relief). Similar to the classical Relief algorithms, the goal of DRNSR-Relief is to maximize the expected margin in the weighted feature space. A dynamic representation framework is introduced to show the dynamic relationship between the expected margin vector and the weight vector. To achieve better neighbor reconstruction, DRNSR-Relief decomposes a nonlinear problem into a set of locally linear ones through local hyperplane with l_1 regularization and then estimates feature weights in a large margin framework. With the help of gradient ascent method, we can guarantee the convergence of DRNSR-Relief. To demonstrate the validity and the effectiveness of our formulation for feature selection in supervised learning, we perform extensive experiments on synthetic and real-world datasets. Experimental results indicate that DRNSR-Relief is very promising.

1. Introduction

Nowadays, datasets are characterized by hundreds or even thousands of features, which may consist of irrelevant noises. Hence, dimensionality reduction is an important task in machine learning and pattern classification [1–4]. Feature selection as a mean of dimensionality reduction in machine learning and pattern recognition has attracted a lot of attention of researchers. Feature selection aims to select the most representative feature subset with a high performance by eliminating redundant and unimportant features. Generally speaking, feature selection has three advantages. First, feature selection can reduce the dimension of given data. Second, feature selection can enhance the generalization performance of classifiers in case of feature redundancy. In other words, classifiers modeled by an optimal feature subset can improve the classification accuracy. Third, feature selection can deepen the understanding of data when data visualization is possible.

The existing feature selection algorithms can be categorized as filter and wrapper methods based on criterion functions which are used in searching informative features [5]. Filter methods are independent of classifiers and can select a feature subset from an original dataset using specific evaluation criteria which are mostly based on interclass distance measurement (e.g., Fisher score) or statistical methods (e.g., p-value and t-test) [6–11]. Wrapper methods employ the performance of classifiers to evaluate the importance of feature subsets and pick up an optimal one. Therefore, filter methods are computationally more efficient than wrapper methods, but usually do not perform as well as wrapper methods. Both filter and wrapper methods require criterion functions for searching informative features. Generally, it is necessary to apply a searching strategy to select features [12–16]. An exhaustive search is optimum, but it quickly becomes computationally infeasible with the increase of problem size. Hence, some heuristic combinational searches (e.g., forward and backward selection [17]) have been proposed to alleviate this issue. Moreover, these algorithms have been successfully applied to practical situations. However, none of them can provide any guarantee of optimality [18].

To solve the computational complexity issue of searching strategies, feature weight estimation, the counterpart to feature selection, has some merits. In contrast to feature selection, the diagonal

* Corresponding author.
E-mail addresses: 20154227006@stu.suda.edu.cn (X. Huang), zhangliml@suda.edu.cn (L. Zhang), wangbangjun@suda.edu.cn (B. Wang), cszzhang@suda.edu.cn (Z. Zhang), lfzh@suda.edu.cn (F. Li).

https://doi.org/10.1016/j.patcog.2018.03.014

elements of the projection matrix in feature weight estimation are allowed to be real-valued numbers instead of binary ones which can be induced by some well-established optimization techniques for simplicity and effectiveness [8,10,11,19–21]. There are two advantages of feature weighting: there is no need to pre-define the number of relevant features, and standard optimization technologies can be employed to avoid combinatorial search [22,23]. Owing to the performance feedback of a nonlinear classifier in search for features, Relief [8] is considered as a typically successful feature weighting algorithm. The main idea behind Relief is to iteratively update feature weights according to their discriminative ability between neighboring patterns. But Relief is only for binary classification tasks. Further, Relief has been extended to Relief-F for multi-class classification tasks [19]. Relief-F uses multiple nearest neighbors instead of just one nearest neighbor when computing the distance margin.

However, the nearest neighbors defined in the original space are highly unlike the ones in the weighted space [24]. Thus, Sun et al. proposed a new algorithm called I-Relief based on the theoretical framework which has been applied to solve the issue of outliers [10]. The margin defined in I-Relief is obtained by averaging the margin of samples with nearest neighbors. Therefore, feature weight estimation may be less accurate if the samples contain abnormal samples or much irrelevant features. To remedy it, Cai et al. proposed a new method which estimates feature weights from local patterns approximated by a locally linear hyperplane, called as LH-Relief [11]. In the Relief's family, the performance of feature selection is largely determined by the neighborhood representation and the weight updating.

This paper proposes a novel feature weight estimation method, dynamic representation and neighbor sparse reconstruction-based Relief (DRNSR-Relief). The goal of DRNSR-Relief is to maximize the expected margin in the weighted feature space. We propose a dynamic representation framework to describe the optimization problem of margin maximization and provide an effective method to solve the optimization problem.

Some popular Relief based feature weighting algorithms, such as Relief, I-Relief and LH-Relief can be unified in the proposed framework. To achieve better neighbor reconstruction, we construct nearest neighbors for a given point using a sparse reconstruction technique with l_1 regularization. We highlight the contributions of this paper as follows:

- We propose a new dynamic representation framework for feature weight estimation, which redefines the optimization problem. The dynamic representation framework reveals the dynamic process of the weight iteration clearly, and shows the dynamic relationship between the expected margin vector and the weight vector. In addition, traditional Relief methods can be redefined and analyzed under this framework.
- Using gradient ascent method, we provide an effective method to solve the optimization problem of DRNSR-Relief and can guarantee its convergence.
- A novel neighbor sparse reconstruction method is proposed for represent neighbors of the given samples, which is an alternative method of the neighbor representation using local hyperplane learning.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work, including Relief, Relief-F, I-Relief and LH-Relief. DRNSR-Relief is presented in Section 3. Section 4 gives extensive experimental results and analyzes the proposed model. Conclusions are provided in Section 5.

2. Related work

In this section, we describe the related work on Relief and its variants.

2.1. Relief

Relief was first proposed for binary classification tasks [8]. The main idea behind Relief is to estimate feature weights iteratively according to their discriminative ability between patterns from different classes. Let the training sample set be $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \in \mathcal{X} \times \mathcal{Y}$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^I$, $y_i \in \mathcal{Y} = \{-1, +1\}$ is the class label of $\mathbf{x}_i$, N and I are the number and the dimension of training samples, respectively. In each iteration, a pattern $\mathbf{x}_i$ is first randomly selected and then its nearest neighbors $\mathbf{x}_i^{NH}$ from the same class (referred as nearest hit or NH) and $\mathbf{x}_i^{NM}$ from the opposite class (referred as nearest miss or NM) are required, respectively. According to $\mathbf{x}_i$ and its nearest neighbors, Relief updates the feature weights:

$$w_j = w_j + |x_{ij} - x_{ij}^{NM}| - |x_{ij} - x_{ij}^{NH}|, j = 1, \dots, I \tag{1}$$

where w_j is the weight for feature j, x_{ij}, x_{ij}^{NM}, and x_{ij}^{NH} are the jth component of $\mathbf{x}_i$, $\mathbf{x}_i^{NM}$ and $\mathbf{x}_i^{NH}$, respectively.

2.2. Relief-F

Relief-F [19] is an extension of Relief for solving the issue that Relief is limited to classification problems with two classes. Similar to Relief, Relief-F randomly selects an instance $\mathbf{x}_i$, but searches its k nearest neighbors from the same class, called nearest hits $\mathbf{x}_i^{NH_m}, m = 1, \dots, k$, and k nearest neighbors from different classes, called nearest misses $\mathbf{x}_i^{NM_m}, m = 1, \dots, k$. Relief-F updates the quality estimation $\mathbf{w}$ for all attributes depending on their values in $\mathbf{x}_i$, hits $\mathbf{x}_i^{NH_m}$ and misses $\mathbf{x}_i^{NM_m}$. The update formula is similar to that of Relief, except that we have to average the contribution of all the hits and all the misses.

$$w_j = w_j - \frac{diff\left(j, \mathbf{x}_i, X_i^{NH}\right)}{k * T} + \frac{diff\left(j, \mathbf{x}_i, X_i^{NM}\right)}{k * T} \tag{2}$$

where X_i^{NH} and X_i^{NM} denote the sets of nearest hits and nearest misses, respectively,

$$diff\left(j, \mathbf{x}_i, X_i^{NH}\right) = \sum_{m=1}^{k} \frac{|x_{ij} - x_{ij}^{NH_m}|}{max(\{x_{.j}\}) - min(\{x_{.j}\})}$$

and

$$diff\left(j, \mathbf{x}_i, X_i^{NM}\right) = \sum_{m=1}^{k} \frac{|x_{ij} - x_{ij}^{NM_m}|}{max(\{x_{.j}\}) - min(\{x_{.j}\})}$$

The contribution for each class of the misses is weighted with the prior probability of that class. The process is repeated for T times.

2.3. I-Relief

Sun et al. proposed I-Relief to overcome the drawbacks of Relief, such as outlier detection and inaccurate updates [10]. I-Relief generalizes the updating scheme to compute the maximum expected margin $\overline{\rho_i}(\mathbf{w})$ by scaling the features. To account for the uncertainty in defining local information, I-Relief uses a probabilistic model where the nearest neighbors of a given sample are treated as latent variables. Following the principles of the expectation-maximization (EM) algorithm [25], $\overline{\rho_i}(\mathbf{w})$ is computed by averaging out the latent variables:

$$\overline{\rho_i}(\mathbf{w}) = \mathbf{w}^T \Big(\sum\nolimits_{n \in M_i} P(\mathbf{x}_n = \mathbf{x}_i^{NM}|\mathbf{w}) |\mathbf{x}_i - \mathbf{x}_n| - \sum\nolimits_{n \in H_i} P(\mathbf{x}_n$$