# Autoencoder Feature Selector

to use neural network to learn flexible nonlinear relationship among features. The powerful

Embedded feature selection methods can combine feature selection with training process into a whole part. The most widely used embeded methods are regularization models that introduce additional constraints into the optimization of a predictive algorithm that bias the model toward lower complexity. One of famous regularization algorithms is the Lasso proposed by Tibshirani in 1996 [22]. Let X be the feature matrix and y the label vector. The

usually in the form of least square or cross entropy loss. In this paper, we just use the least square loss as the fitting error: $\mathcal{J}_{AE}\left(X, g(f(X))\right) = \frac{1}{2m}\sum_{i=1}^{m} \|\boldsymbol{x}_i - g(f(\boldsymbol{x}_i))\|_2^2 = \frac{1}{2m}\|X - g(f(X))\|_F^2$.

## 3.2 The Proposed Model

the $i$th feature contributes little to the representation of other features; on the other hand, if $i$th feature plays important role in the representation of other features, then $\|\boldsymbol{w}_i\|_2$ must be significant. To select the most discriminative features from original ones, we impose rowsparse regularization on $W^{(1)}$. That is to say, we use $\mathcal{R}(\Theta) = \|W^{(1)}\|_{2,1} = \sum_i^d \sqrt{\sum_j^h (W_{ij}^{(1)})^2}$

$$\min_{W^{(1)},W^{(2)}} \frac{1}{2m}\|X - XW^{(1)}W^{(2)}\|_F^2 + \alpha\|W^{(1)}\|_{2,1}. \qquad (6)$$

We can find that this form is equivalent to (5), so AEFS is a nonlinear extension of RSR.

# 4 Optimization

We construct a synthetic dataset as shown in Fig. 1. The dataset consists of 200 samples with 3 features {A, B, C}. The dataset are split into 2 classes about fifty-fifty. Feature A

**Algorithm 1** Optimization Algorithm of Autoencoder Feature Selector

[6] http://www.ncbi.nlm.nih.gov/pubmed/12381711

[7] http://yann.lecun.com/exdb/mnist/

[8] http://attributes.kyb.tuebingen.mpg.de/

Table 1: Summary of used datasets.

| Dataset | Keywords | #Instances | #Features | #Cl |
|---|---|---|---|---|

[9] Since the result of $k$-means depends on initialization, we repeat the experiments 20 times with random initialization and report the average results with standard deviation.

Given two variables $P$ and $Q$, NMI between them is defined as

| | | | | | | |
|---|---|---|---|---|---|---|
| AWA | 14.4±0.3 | 11.6±0.3 | 13.2±0.2 | 12.2±0.3 | 12.2±0.3 | **13.4±0.4** |

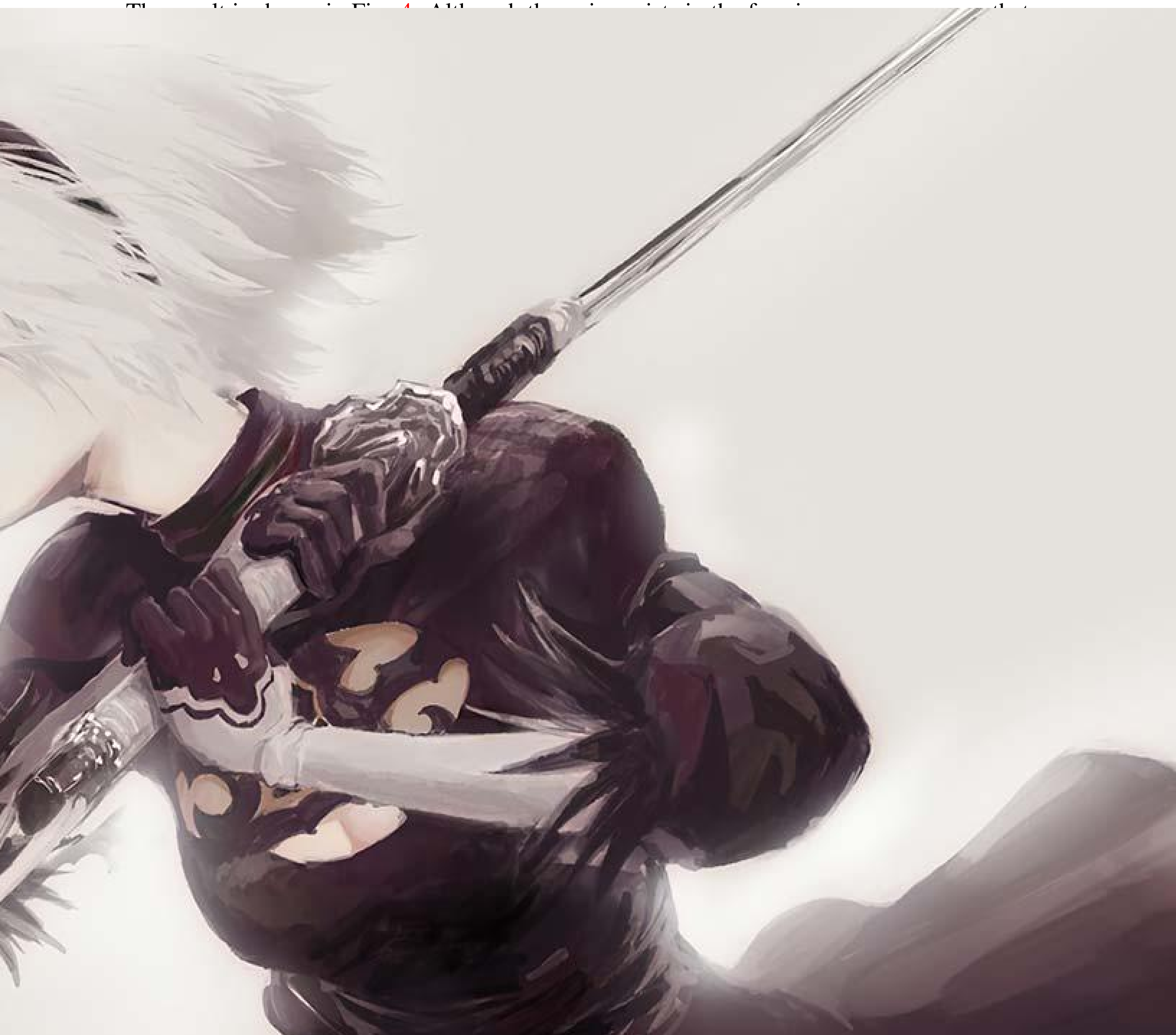
Table 4: Classification results (ACC%) of different feature selection methods. The best results are highlighted in bold.

with much fewer features than the original. However, the reconstructed face of RSR is less similar to raw face than AEFS, especially when the number of selected features is small.

We also evaluate reconstruction ability of denoising AEFS on the face dataset warp-

features.

PIE10P with Gaussian noise. The noise is set zero mean value and 0.5 standard deviation. The result is shown in Fig. 4. Although the noise exists in the face images, we can see that



[9] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length, and helmholtz free energy. *Advances in neural information processing systems*, pages 3–3, 1994.

[10] Emanuel F. Petricoin Iii, David K. Ornstein, Cloud P. Paweletz, Ali Ardekani, Paul S. Hackett, Ben A. Hitt, Alfredo Velassco, Christian Trucco, Laura Wiegand, and Kamil

[24] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[25] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. l2, 1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*