Unsupervised Feature Selection using Pseudo Label Approximation

REN DENG
Amazingx Academy, China

YE LIU
Amazingx Academy, China

LIYAN LUO
Amazingx Academy, China

DONGJING CHEN
Amazingx Academy, China

XIJIE LI*
Xi'an Institute of Optics and Precision
Mechanics, Chinese Academy of
Sciences, Xian, China

ABSTRACT

Feature selection is a machine learning technique that selects a representative subset of all features available in order to reduce the time and space needed to process high-dimensional data. Traditional feature selection methods include filter, wrapper, and embedded approaches. However, many conventional methods' performances are not suitable in many contexts. This paper proposes a new unsupervised feature selection model based on pseudo label approximation. The new derived model incorporates a projection error, a sparsity regularization, and a manifold regularization term that preserves the manifold structure of the original data. Finally, implementation of the new model onto five distinct datasets validates the effectiveness of the proposed model.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Machine learning; • **Theory of computation** → Models of computation.

KEYWORDS

Feature selection, Pseudo label matrix, High-dimensional data

ACM Reference Format:

REN DENG, YE LIU, LIYAN LUO, DONGJING CHEN, and XIJIE LI. 2021. Unsupervised Feature Selection using Pseudo Label Approximation. In *2021 13th International Conference on Machine Learning and Computing (ICMLC '21)*, February 26–March 01, 2021, Shenzhen, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3457682.3457758>

1 INTRODUCTION

In the modern age of Big Data, high-dimensional data is increasingly utilized in a wide variety of fields, including predictive analytics, computer vision, and artificial intelligence. High-dimensional data can usually offer a more detailed spectrum of information that may be valuable to data analytics; however, it is often inefficient and

time-consuming to process high-dimensional data with the limited amount of storage available. Moreover, high-dimensional data often contain redundant or irrelevant features that are not significant for the analytical task. In order to address this issue, feature selection, or the careful choosing of a representative subset of the original features, has become increasingly important in machine learning as a necessary preprocessing step [1].

Feature selection can significantly improve a system's efficiency and efficacy by removing the redundant features—features that can be derived from other features—and the irrelevant features—features that are not related to the problem. The method can be primarily categorized into two types: (1) supervised feature selection and (2) unsupervised feature selection. Supervised feature selection analyzes and selects features that are labeled, meaning that they possess intrinsic information that can evaluate their performances. Unsupervised feature selection, on the other hand, deals with information that is unlabeled. While unlabeled feature selection is usually more challenging due to the lack of intrinsic label that can evaluate the performance, it is still an important task, as there are far more unlabeled data in real life than labeled, structured data [5]. Considering the significance of unsupervised learning problems, this paper focuses primarily on the development of unsupervised feature selection.

Traditionally, unsupervised feature selection approaches can be primarily categorized into (1) filter methods, (2) wrapper methods, and (3) embedded methods. The filter methods are the most straightforward approaches that select a subset of features based on their statistical properties such as variance. These methods do not employ any learning algorithms to select the features [1–3]. They focus on the intrinsic statistical properties of the algorithm and select features based on them. For example, a filter method may select the features that have the largest variance, which are assumed to be more representative than those with smaller variance. Despite the seemingly effective and straightforward mechanism, this approach often overlooks the presence of redundant features. This inability to recognize redundancy in the feature set thus renders the method susceptible to inaccurate performance. Consequently, it could fail to select the most relevant features in many important tasks. The wrapper methods, on the other hand, evaluate the features based on a certain learning algorithm [4, 5, 6]. These methods center around a predictive analytical model to determine the most relevant features. They are often employed for data clustering and feature selection process, measuring the relevancy of features based on their performance in a subset. As a matter of fact, this

* Xijie Li is the corresponding author. His email is lixijie@opt.ac.cn, respectively.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMLC '21, February 26–March 01, 2021, Shenzhen, China

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8931-0/21/02...\$15.00

<https://doi.org/10.1145/3457682.3457758>

type of feature selection methods is often more effective than its filter counterparts; nevertheless, the approach is sometimes overly time consuming for feature selection process, so more sophisticated algorithm is needed [1]. More recently, the embedded methods for feature selection become more popular [7], as they appear superior in performance compared with others. These methods carry out the feature selection process during model training phase; in other words, this type of methods is incorporated into the model development phase, and it can identify numerous statistical properties of the data.

In this paper, we propose an unsupervised feature selection method by leveraging pseudo label approximation. The model includes a projection error term, a sparsity regularization constraint, and a term preserving the manifold structure of the data. Moreover, an alternative iterative algorithm is used to optimize the model function.

2 PSEUDO LABEL APPROXIMATION

This paper proposes a novel unsupervised feature selection model utilizing pseudo label approximation. While unsupervised feature selection methods are challenging to tackle, generating pseudo labels for the features could effectively change the problem into a supervised learning problem. To begin with, let $X \in R^{d \times n} = [x_1, \dots, x_d]$ denote a d by n matrix dataset, where d is the number of samples and n is the number of features. $x_i \in R^{d \times 1}$ denotes the i -th row of X .

When X is projected to the pseudo label space using linear projection, its projection error can be represented as:

$$\|X^T W + I_n b^T - Y_p\|_F^2, \text{ s.t. } Y_p^T Y_p = I_c,$$

where $W \in R^{d \times C}$ denotes the projected matrix, $b \in R^{C \times 1}$ is bias, Y_p represents pseudo label matrix, I_n represents the n -dimension all-1 vector, and I_c represents a $C \times C$ identity matrix.

Because feature selection is sparsity problem, sparsity regularization should be added. The sparsity regularization is written as follows:

$$\alpha \|W\|_{2,1},$$

where α is hyper-parameter (a constant), and $\|\cdot\|_{2,1}$ represents 2, 1-norm. W^i denotes the i -th row vector of W , and $\|W\|_{2,1}$ can be represented as follows:

$$\|W\|_{2,1} = \sum_{i=1}^d \|W^i\|_2,$$

where $\|W^i\|_2$ represents the length of W^i .

In this paper, manifold learning is also used to preserve the local structure of the original data. Specifically, the pseudo label matrix can be seen as the low dimensional manifold of the original data, and the manifold regular can be represented as follows

$$\beta Y_p^T L Y_p,$$

where β is hyper-parameter and L is the Laplacian matrix. The Laplacian matrix can be calculated as follows:

$$L = D - S,$$

where $S \in R^{n \times n}$ is the data similarity matrix and D is a diagonal matrix, and the diagonal elements can be computed as follows:

$$d_{ii} = \sum_{j=1}^n S_{ij},$$

where $S = [S_{ij}]_{n \times n}$ is computed as follows:

$$S_{ij} = \frac{\exp\left\{-\frac{\|x_i - x_j\|_2^2}{\sigma^2}\right\}}{\sum_{i=1}^n \exp\left\{-\frac{\|x_i - x_j\|_2^2}{\sigma^2}\right\}},$$

For this end, the whole feature selection model proposed by this paper can be summarized as follows:

$$\begin{aligned} \min_{W, b, Y_p} & \|X^T W + I_n b^T - Y_p\|_F^2 + \alpha \|W\|_{2,1} + \beta Y_p^T L Y_p \\ \text{s.t. } & Y_p^T Y_p = I_c, \sum_{i=1}^n S_{ij} = 1, 0 < S_{ij} < 1 \end{aligned},$$

2.1 Optimization of Model

This subsection attempts to optimize the above feature selection model. First of all, to solve for b , we need to solve for the partial derivative of b in the above equation, and set that partial derivative to zero. This obtains the following expression for b :

$$b = \frac{1}{n} (Y_p^T - W^T X) \mathbf{1}_n,$$

Substitute the above expression for b back to Eq. 1 and simplify:

$$\begin{aligned} \min_{W, Y_p} & \|C_n (X^T W - Y_p)\|_F^2 + \alpha \|W\|_{2,1} + \beta Y_p^T L Y_p, \text{ s.t. } Y_p^T Y_p \\ & = I_c, \sum_{i=1}^n S_{ij} = 1, 0 < S_{ij} < 1, \end{aligned}$$

where $C_n = I_n - \frac{1}{n} \mathbf{1}_n \cdot \mathbf{1}_n^T$.

However, while b can be directly solved for, others must be solved using an alternative iterative algorithm, as it is hard to solve for the optimal solution directly. An alternative iterative algorithm needs to solve for W and Y_p separately, each time setting the other variable to be constant.

2.1.1 Algorithm one: Optimization of W . When Y_p is fixed, the above equation can be rewritten as the following:

$$\min_W \|C_n (X^T W - Y_p)\|_F^2 + \alpha \|W\|_{2,1} \triangleq L_1(W).$$

When the partial differential of W is set to equal to 0, we can derive the following equation:

$$\frac{\partial L_1(W)}{\partial W} = 2XC_n X^T W - 2XC_n Y_p + 2\alpha QW = 0.$$

According to the above equation, the updating rule of W can be written as follows:

$$W \leftarrow (XC_n X^T + \beta Q)^{-1} XC_n Y_p.$$

Therefore, the pseudocode for the alternative iterative algorithm for W can be represented as:

Algorithm 1: alternative iterative algorithm for W **Input :** $X \in R^{d \times n}, Y_p, \alpha$ **Initialization :** $Q = I_d$ **Output :** W **Repeat:**

$$W \leftarrow (XC_n X^T + \beta Q)^{-1} XC_n Y_p$$

$$Q = \begin{bmatrix} \frac{1}{2 \|W^1\|} & & & \\ & \frac{1}{2 \|W^2\|} & & \\ & & \ddots & \\ & & & \frac{1}{2 \|W^d\|} \end{bmatrix}$$

until convergence

2.1.2 Algorithm two: Optimization of Y_p . Next, consider the case when W is fixed. the above equation will then be equivalent to:

$$L(Y_p) = \|C_n(X^T W - Y_p)\|_F^2 + \beta Y_p^T L Y_p, \text{ s.t. } Y_p^T Y_p = I.$$

The above equation is ultimately equivalent to the following equation:

$$\begin{aligned} L'(Y_p) &= \text{tr}(Y_p^T (C_n + \beta L) Y_p) - 2 \text{tr}(Y_p^T C_n X^T W) \\ \text{s.t. } Y_p^T Y_p &= I_c \end{aligned}$$

The optimization algorithm for the above problem can thus be illustrated as follows:

Algorithm 2: alternative iterative algorithm for Y_p **Input :** $X \in R^{d \times n}, L \in R^{n \times n}, W \in R^{n \times d}, \beta > 0$ **Output :** $Y_p \in R^{n \times c}$ **Initialization :** $Y_p^T \cdot Y_p = I_c$ **Calculate:** $A = C_n + \beta L, A' = \lambda_A I_n - A$ λ_A represents A 's largest eigenvalue $B = C_n X^T W$ **Repeat**1. Update $M \leftarrow A' Y_p + 2B$ 2. Let $M = U_M \sum V_M^T$ (singular value decomposition)3. Update $Y_p = U_M V_M^T$ **until convergence**

2.1.3 Overall Algorithm. Overall, the feature selection algorithm proposed by this paper can be summarized as follows:

Algorithm 3: Alternative iterative algorithm to solve problem**Input :** $X \in R^{d \times n}, L \in R^{n \times n}, W \in R^{n \times d}, \alpha > 0, \beta > 0$ **Output :** N selected features.**Initialize** $Y_p \in R^{n \times c}$, s.t. $Y_p^T Y_p = I_c$, $\Lambda = I_d$ **Calculate :** S, D and L **Repeat:**Update W , using algorithm 1Update Y_p , using algorithm 2Calculate D and L **until convergence**

Select K features: rank the features according to the values of $\sum_{j=1}^d W_{ij}^2$ in descending order;

choose the first K features corresponding to the first K values' corresponding i values

3 EXPERIMENTAL RESULTS

This section verifies the effectiveness of the proposed model on five publicly available datasets regarding alphabet, image data, and biological data. The model is evaluated on three image datasets (USPS, MNIST, COIL20), one biological dataset (TOX_171), and one alphabet dataset (Isolet). Information about the datasets used

are shown by Table 1 below. Clustering accuracy (AC) is utilized as the metric for performance evaluation [12–14], as it has been used in previous work on unsupervised feature selection as well [8]. To compare the new model's performance with other alternative methods, this paper also carries out several other common methods for feature selection, including Laplacian score (LS), Multi-Cluster Feature Selection (MCFS), Unsupervised Discriminative Feature Selection (UDFS), and Nonnegative Discriminative Feature Selection (NDFS). In addition, the experiment is also run with all existing features (All-Fea) to demonstrate how feature selection may affect clustering performance. All original features are utilized in that case. The mechanisms of these feature selection methods are briefly summarized below:

1. LS selects features based on their locality preserving ability and variance, and it is independent of any learning algorithm [10];
2. MCFS selects features based on their multi-cluster structure preserving ability, though it fails to make use of the discriminative information in feature selection process [5];
3. UDFS selects features using a model that incorporates discriminative analysis with $l_{2,1}$ -norm minimization regularization [9];
4. NDFS carries out cluster label learning and feature selection at the same time, with a nonnegative constraint term to improve accuracy. This method exploits the discriminative information and feature correlation simultaneously to select the most discriminative features [11].

The convergence curves of the proposed model applied on the datasets and the experimental results are shown below in Figure 1 and Table 2

Two major conclusions can be drawn from the experimental results. First, it can be observed that unsupervised feature selection methods often (though not always) work better than the method that employs all features available. This demonstrates that feature selection is indeed useful and valuable, as it not only increases the algorithm's efficiency by reducing run-time, but also improves the overall performance. Second, the proposed model consistently outperforms traditional feature selection methods. For example, the proposed method's accuracy on the MNIST dataset has more than 4% accuracy than the second best method used (MCFS). It can therefore be inferred that the new model is more effective in selecting a representative subset of features.

4 CONCLUSION

In conclusion, this paper develops an unsupervised feature selection method that is based on pseudo label approximation. The model includes a projection error term, a sparsity regularization constraint, and a term preserving the manifold structure of the data. Moreover, an alternative iterative algorithm is used to optimize the model function. Implementation of the new model onto five real-world applications also validates the model's accuracy, rendering it more accurate and effective than the preexisting feature selection methods.

ACKNOWLEDGMENTS

This work was conducted and supported under the guidance of Ms. Fleur Gooden from BASIS International School Guangzhou. The

Figure 1: Convergence curves of the proposed model on tested datasets.

Table 1: Dataset information

Dataset	Number of samples	Number of features	Number of categories	Dataset type
Isolet	1560	617	26	Letters
USPS	9298	256	10	Image
COIL20	1440	1024	20	Image
MINIST	3000	784	10	Image
TOX_171	171	5748	4	Biological Data

Table 2: Comparison of clustering accuracy (ACC, %). The best results for each dataset are shown in bold.

Feature selection methods	Isolet	USPS	COIL20	MINIST	TOX_171
All-Fea	64.08	68.40	54.20	41.01	41.78
LS	56.92	70.19	60.88	40.68	45.72
MCFS	60.27	71.97	60.93	56.19	46.97
UDFS	48.83	68.78	61.45	44.97	36.53
NDFS	68.34	67.78	60.99	42.59	42.97
Ours	69.87	73.68	64.33	60.87	49.22

research model was independently proposed and developed by Ren Deng; it was originally inspired by Quanmao Lu et al.'s paper titled "Structure preserving unsupervised feature selection." I would like to give special thanks to all the references listed above, which have helped me tremendously throughout my research and modeling process.

REFERENCES

- [1] Yu, Lei, and Liu, Huan. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution." *aaai.org/Papers/ICML/2003/ICML03-111.pdf*.
- [2] Sánchez-Marono, Noelia, *et al.* "Filter methods for feature selection. A comparative study." https://www.researchgate.net/profile/Amparo_AlonsoBetanzos/publication/221252792_Filter_Methods_for_Feature_Selection_A_Comparative_Study/links/543fd9ec0cf21227a11b8e05.pdf.
- [3] Talavera, Luis. "An evaluation of filter and wrapper methods for feature selection in categorical clustering." https://www.researchgate.net/profile/Luis_Talavera/publication/221460675_An_Evaluation_of_Filter_and_Wrapper_Methods_for_Feature_Selection_in_Categorical_Clustering/links/0deec53aec6c5a80fc000000/An-Evaluation-of-Filter-and-Wrapper-Methods-for-Feature-Selection-in-Categorical-Clustering.pdf.

- [4] Jović, A., *et al.* "A review of feature selection methods with applications." https://bib.irb.hr/datoteka/763354.MIPRO_2015_JovicBrkicBogunovic.pdf.
- [5] Lu, Quanmao, *et al.* "Structure preserving unsupervised feature selection." August 2018. <https://doi.org/10.1016/j.neucom.2018.04.001>.
- [6] Wah, Yap Bee, *et al.* "Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximising Classification Accuracy." https://www.researchgate.net/profile/Shuzlina_Rahman/publication/322920304_Feature_selection_methods_Case_of_filter_and_wrapper_approaches_for_maximising_classification_accuracy/links/5eb17fa545851592d6b9b50c/Feature-selection-methods-Case-of-filter-and-wrapper-approaches-for-maximising-classification-accuracy.pdf.
- [7] Huang, Suhan, *et al.* "Embedded Unsupervised Feature Selection." https://www.researchgate.net/profile/Suhang_Wang/publication/308891942_Embedded_Unsupervised_Feature_Selection/links/57f53d0508ae8da3ce5521ac/Embedded-Unsupervised-Feature-Selection.pdf.
- [8] Cao, Xiaochun, *et al.* "Diversity-induced Multi-view Subspace Clustering." Computer Vision Foundation. https://openaccess.thecvf.com/content_cvpr_2015/papers/Cao_Diversity-Induced_Multi-View_Subspace_2015_CVPR_paper.pdf.
- [9] Yang, Yi, *et al.* "L2,1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning." <https://opus.lib.uts.edu.au/bitstream/10453/119490/1/267.pdf>.
- [10] He, Xiaofei, *et al.* "Laplacian Score for Feature Selection." <http://papers.nips.cc/paper/2909-laplacian-score-for-feature-selection.pdf>.
- [11] Li, Zechao, *et al.* "Unsupervised Feature Selection Using Nonnegative Spectral Analysis." <http://nlpr-web.ia.ac.cn/2012papers/gjhy/gh27.pdf>.
- [12] Yaxiong Chen, Xiaoqiang Lu, and Xuelong Li. Supervised deep hashing with a joint deep network. Pattern Recognition. 2020, 105. 107368. doi: 10.1016/j.patcog.2020.107368.
- [13] Y. Yuan, H. Ning, and X. Lu, "Bio-Inspired Representation Learning for Visual Attention Prediction," IEEE Transactions on Cybernetics, 2019.
- [14] H. Ning, X. Zheng, Y. Yuan, and X. Lu, "Audio Description from Image by Modal Translation Network," Neurocomputing, 2020.