Feature Selection for Clustering by Exploring Nearest and Farthest Neighbors

Chien-Hsing Chen
Department of Information Management, Hwa Hsia Institute of Technology
Taipei, Taiwan
ktfive@gmail.com

***Abstract*—Feature selection has been explored extensively for use in several real-world applications. In this paper, we propose a new method to select a salient subset of features from unlabeled data, and the selected features are then adaptively used to identify natural clusters in the cluster analysis. Unlike previous methods that select salient features for clustering, our method does not require a predetermined clustering algorithm to identify salient features, and our method potentially ignores noisy features, allowing improved identification of salient features. Our feature selection method is motivated by a basic characteristic of clustering: a data instance usually belongs to the same cluster as its geometrically nearest neighbors and belongs to a cluster different than those of its geometrically farthest neighbors. In particular, our method uses instance-based learning to quantify features in the context of the nearest and the farthest neighbors of every instance so that clusters generated by the salient features maintain this characteristic.**

Keywords-feature selection; nearest neighbor; farthest neighbor; salient feature; cluster analysis.

I. INTRODUCTION

Feature selection has been explored extensively for use with several real-world applications, such as text processing [1] and image representation [2]. Typically, a larger number of features to represent the patterns is more informative for a learning algorithm. However, in a high-dimensional dataset, some features are noisy, and thus the learning algorithms often suffer from the bias of noisy features that influence the learning process. Recently, many extensive studies have been proposed for feature selection with unsupervised learning, and the selected salient feature subsets were found to aid cluster analysis [3-6]. The goal of feature selection for clustering is to identify a subset of relevant features and remove redundancy from the original representation space. In addition, feature selection is also used to choose selected features to partition a dataset into clusters while effectively increasing both the cluster compactness for the data instances within a cluster and the cluster separability of the data instances between clusters.

In this paper, we propose a new method for selecting a subset of original features from unlabeled data, and the selected feature subset is adaptively used to identify natural clusters in the cluster analysis. The main contribution of this work is that our method achieves the goal of feature selection for clustering without the need to exactly explore the clustered information, thus potentially ignoring the bias of noisy or uninformative features that influence the identification of salient features.

Our method uses instance-based learning for quantifying features in the context of the nearest and the farthest neighbors of every instance. This quantification is motivated by one of the most well-known characteristics of clustering: an instance usually belongs to the same cluster as its geometrically nearest neighbors and belongs to a cluster different than those of its geometrically farthest ones. Therefore, the purpose of our feature selection method is to quantify features so that clusters generated by the salient features (i.e., with higher quantity) maintain this well-known characteristic. Therefore, our method is advantageous because our method does not need to explore natural clusters using a predetermined clustering algorithm.

With our method, a feature is quantified by its ability to distinguish between the nearest and the farthest neighbors of every instance. The quantifying features learning process is to identify the best feature salience vector (represented by a real-valued quantity vector to indicate its salience for this distinguishability) instead of to heuristically search a subset of features in the space of all possible feature subsets. In addition, we implement a gradient descent iterative method employing cooperative and competitive strategies to identify the best feature salience vector.

II. THE PROPOSED METHOD

2.1 Observations of our Proposed Method

We present an example in which we have a set of instances each with two dimensions, "Feature 1" and "Feature 2" (Fig. 1), and all instances can be clustered into two assumptive clusters, "Cluster 1" and "Cluster 2." For this example, we discuss our presented feature selection method to extract salient features which are adaptively used for discovering natural clusters.

We briefly introduce our method for determining a salient feature. This method begins with the instance "$\mathbf{x}_1$", which has both specific nearest and farthest neighbors (see Figure 1). We define feature separability as the average distance from an instance to its farthest neighbors (its magnitude is represented as a dotted line) and the feature compactness as the average distance from this instance to its nearest neighbors (its magnitude is represented as a solid line), where both distances are measured with respect to a particular feature. We assume

 ISBN: 978-1-61208-217-2

that a feature is more salient if it yields a higher value of the following measure:

$$\frac{\#\{\text{average length of corresponding to the separability}\}}{\#\{\text{average length of corresponding to the compactness}\}}$$

Therefore, based on this assumption, "Feature 1" is more salient. In the context of clustering, for which we want to partition the dataset into clusters, we would be much more likely to believe that the contribution of "Feature 1" is higher than that of "Feature 2."

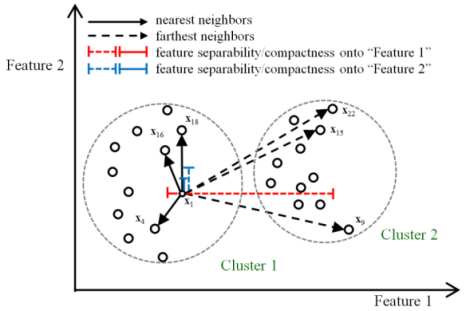


Figure 1. A schematic example: an instance "$\mathbf{x}_1$" has its nearest neighbors {"$\mathbf{x}_{16}$", "$\mathbf{x}_{18}$", "$\mathbf{x}_4$"} and its farthest neighbors {"$\mathbf{x}_9$", "$\mathbf{x}_{22}$", "$\mathbf{x}_{15}$"}. For the instance "$\mathbf{x}_1$", "Feature 1" should be more salient than "Feature 2".

2.2 Preliminaries of our Proposed Method

In this paper, we present a new unsupervised feature selection method that does not require any clustering learning algorithm to identify salient features; the salient features selected by our method can be then effective for clustering. Our method is based on a basic characteristic of clustering: an instance usually belongs to the same cluster as do its nearest neighbors and belongs to a cluster different than those of its farthest neighbors. With our method, a feature is quantified by its ability to distinguish between the nearest and the farthest neighbors of every instance.

Assume a dataset X of n data instances ($\boldsymbol{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$), where $\mathbf{x}_i = [x_{1,i}, \ldots, x_{j,i}, \ldots, x_{d,i}]^{\mathrm{T}}$ represents the i^{th} instance in $\boldsymbol{X}$ with d dimensions; also assume a non-zero feature salience vector $\mathbf{w}(t) = [w(t)_1, \ldots, w(t)_j, \ldots, w(t)_d]^{\mathrm{T}}$, where the element $w(t)_j$ is a real-valued quantity at the t^{th} iteration. We first consider $\mathbf{w}(t)$ to obtain the nearest and the farthest neighbors for a given instance. The k^{th} nearest neighbor $\mathbf{x}^{\Theta}_{i\to k,\mathbf{w}(t)}$ and the l^{th} farthest neighbor $\mathbf{x}^{\phi}_{i\to l,\mathbf{w}(t)}$ of $\mathbf{x}_i$ are subject to, respectively,

$$\pi(k) = \sum_{r=1, r\neq i}^{n} I\Big(dist\big(\mathbf{x}_i, \mathbf{x}_r|\mathbf{w}(t)\big) \leq dist\Big(\mathbf{x}_i, \mathbf{x}^{\Theta}_{i\to k;\,\mathbf{w}(t)}|\mathbf{w}(t)\Big)\Big) \quad (1)$$

$$\pi(l) = \sum_{r=1, r\neq i}^{n} I\Big(dist\big(\mathbf{x}_i, \mathbf{x}_r|\mathbf{w}(t)\big) \geq dist\Big(\mathbf{x}_i, \mathbf{x}^{\phi}_{i\to k;\,\mathbf{w}(t)}|\mathbf{w}(t)\Big)\Big) \quad (2)$$

where $\pi()$ transfers an ordinal number to an interval number that represents the number of instances that satisfy the whole condition. Θ represents a nearest neighbor and ϕ represents a farthest neighbor. $I()$ outputs 1 when the condition is satisfied and outputs zero otherwise. $dist(\mathbf{x}_a,\mathbf{x}_b|\mathbf{w}(t))$ is a distance function in which we use the weighted Euclidean metric to measure the distance between $\mathbf{x}_a$ and $\mathbf{x}_b$ under $\mathbf{w}(t)$.

$$dist(\mathbf{x}_a, \mathbf{x}_b \mid \mathbf{w}(t)) = \sqrt{\sum_{j=1}^{d} w(t)_j \times \big(x_{j,a} - x_{j,b}\big)^2} \quad (3)$$

2.2.1 Measure Manhattan Distance with Element-Wise Absolute Operator

We then scale each feature and define two sets of distances from $\mathbf{x}_i$ to its K nearest neighbors and L farthest neighbors. Then, we obtain two new sets of instances, $\mathbf{NS}^{\mathbf{w}(t)}_{i,K}$ and $\mathbf{FS}^{\mathbf{w}(t)}_{i,L}$, which include K and L instances, respectively.

$$\mathbf{NS}^{\mathbf{w}(t)}_{i,K} = \{d\big(\mathbf{x}_i, \mathbf{x}^{\Theta}_{i\to 1;\,\mathbf{w}(t)}\big), \ldots, d\big(\mathbf{x}_i, \mathbf{x}^{\Theta}_{i\to K;\,\mathbf{w}(t)}\big)\} \quad (4)$$

$$\mathbf{FS}^{\mathbf{w}(t)}_{i,L} = \{d(\mathbf{x}_i, \mathbf{x}^{\phi}_{i\to 1;\,\mathbf{w}(t)}), \ldots, d(\mathbf{x}_i, \mathbf{x}^{\phi}_{i\to L;\,\mathbf{w}(t)})\} \quad (5)$$

where $d(.,.)$ is an element-wise absolute operator, thus yielding d-dimensional data.

Let us assume that two sets of neighbors can be distinguished maximally by a subset of features. Therefore, the next step is to extract the salient features. We assign two labels, one for the nearest neighbors and the other for the farthest neighbors. The idea is that our method scales each feature and measures which feature can better achieve separability between the instances in $\mathbf{NS}^{\mathbf{w}(t)}_{i,K}$ and $\mathbf{FS}^{\mathbf{w}(t)}_{i,L}$. First, we define a data fraction $\boldsymbol{\rho}_i(\mathbf{w}(t))$ that includes the instances in $\mathbf{NS}^{\mathbf{w}(t)}_{i,K}$ and $\mathbf{FS}^{\mathbf{w}(t)}_{i,L}$ for a given $\mathbf{x}_i$ under $\mathbf{w}(t)$. The fraction $\boldsymbol{\rho}_i(\mathbf{w}(t))$ is expressed as

$$\boldsymbol{\rho}_i(\mathbf{w}(t)) = \left\{\mathbf{NS}^{\mathbf{w}(t)}_{i,K}, \mathbf{FS}^{\mathbf{w}(t)}_{i,L}\right\} \quad (6)$$

This fraction consists of $K + L$ data instances with d dimensions (i.e., $s_1, \ldots, s_j, \ldots, s_d$, where s_j is the j^{th} variable in

 ISBN: 978-1-61208-217-2

$\boldsymbol{\rho}_i(\mathbf{w}(t))$). Next, we assign a categorical variable c to represent labels for these instances. The label c_r for an instance $\mathbf{x}_r \in \boldsymbol{\rho}_i(\mathbf{w}(t))$ is assigned by

$$c_r = \begin{cases} 0, \text{if } \mathbf{x}_r \in \mathbf{NS}_{i,K}^{\mathbf{w}(t)} \\ 1, \text{if } \mathbf{x}_r \in \mathbf{FS}_{i,L}^{\mathbf{w}(t)} \end{cases} \quad (7)$$

Our method is an unsupervised feature selection method; therefore, the class labels are not considered by the process of feature selection. Fortunately, while the new labels are assigned using Eq. (7), the filter-based and wrapper-based feature selection methods in supervised learning can be used to help our method identify salient features. For example, we can use a filter-based feature selection method (e.g., mutual information) to evaluate how an individual feature informs the target variable [7-9] or apply a wrapper-based feature selection method (e.g., by using SVM to construct a classifier) to observe how a feature better distinguishes between the instances in $\mathbf{NS}_{i,K}^{\mathbf{w}(t)}$ and $\mathbf{FS}_{i,L}^{\mathbf{w}(t)}$ using this classifier [9-12].

2.2.2 Evaluate salient feature using dependency and redundancy metrics

In this paper, we use the filter-based feature selection method to evaluate features because of its efficiency. Thus, we avoid the process of training instances required by the wrapper-based method. The basis of the method to evaluate feature salience is to determine whether a feature is able to distinguish the nearest and the farthest neighbors. In particular, a feature that is more dependent on the target variable and is less redundant with other features is more salient [8]. The criterion to quantify a feature is

$$\Phi_j = D(s_j, c) - R(s_j) \quad (8)$$

in which we use the $D()$ and $R()$ criteria to evaluate dependency and redundancy, respectively. The functions $D()$ and $R()$ are respectively expressed as

$$D(s_j, c) = I(s_j; c) \quad (9)$$

$$R(s_j) = \frac{1}{d-1} \sum_{u=1, u \neq j}^{d} I(s_j; s_{u \neq j}) \quad (10)$$

where $I()$ is a mutual information criterion; other standard criteria can also be used. Because we have a data fraction $\boldsymbol{\rho}_i(\mathbf{w}(t))$ and a categorical variable c, we can obtain a quantification vector $\mathbf{u}_i^{\mathbf{w}(t)} = [u_{1,i}^{\mathbf{w}(t)}, \ldots, u_{j,i}^{\mathbf{w}(t)}, \ldots, u_{d,i}^{\mathbf{w}(t)}]^{\mathrm{T}}$, where $u_{j,i}^{\mathbf{w}(t)}$ represents a quantity measured by Φ_j.

2.3 Searching the Best Feature Salience Vector

In this section, we attempt to search the best feature salience vector $\mathbf{w}$ for which the goal is to satisfy the condition that the $\{\mathbf{u}_i^{\mathbf{w}}\}_{i=1}^{n}$ values are constant for the particular instances $\{\mathbf{x}_i\}_{i=1}^{n} \in X$. Recall that $\mathbf{w}(t)$ is used to find $\mathbf{NS}_{i,K}^{\mathbf{w}(t)}$ and $\mathbf{FS}_{i,L}^{\mathbf{w}(t)}$ applied to evaluate quantification vectors $\{\mathbf{u}_i^{\mathbf{w}(t)}\}_{i=1}^{n}$, which are further applied to reflect $\mathbf{w}(t)$. The criterion to determine the best $\mathbf{w}$ is an NP problem.

We reduce the searching problem to optimize the sum-of-squared error (i.e., $||\mathbf{u}_t^{\mathbf{w}} - \mathbf{w}(t)||^2$). The goal of the learning task is to identify the best $\mathbf{w}$ such that the error is minimized. Therefore, we should use a method will optimize both $\mathbf{u}_t^{\mathbf{w}}$ and $\mathbf{w}(t)$.

We implement a gradient-descent-based approach, Least-Mean-Squares (LMS) [13, 14] with some modifications, and used cooperative and competitive iterative strategies to find $\boldsymbol{W} = [\mathbf{w}(1), \ldots, \mathbf{w}(t), \ldots, \mathbf{w}(T)]^T$. Each iteration t has only one instance $\mathbf{x}_t$ that was randomly selected and participated in learning. Cooperatively, $\mathbf{x}_t$ considers all elements in $\mathbf{w}(t)$ to yield $\mathbf{u}_t^{\mathbf{w}(t)}$. Competitively, $\mathbf{u}_t^{\mathbf{w}(t)}$ is used to perceive the more salient feature and thus inform $\mathbf{w}(t+1)$. Furthermore, $\alpha(t)$ is a monotonically decreasing learning coefficient, so the updated function for $\mathbf{w}(t+1)$ was adaptively written as follows (11). Iterations then stop when $\mathbf{u}_T^{\mathbf{w}}$ and $\mathbf{w}(T)$ are balanced.

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha(t) \times \left[\mathbf{u}_t^{\mathbf{w}(t)} - \mathbf{w}(t)\right] \quad (11)$$

III. EXPERIMENT

This section presents evaluations of our presented method for feature selection on clustering problems. To demonstrate the effectiveness of the selected feature subset for clustering, the usefulness is evaluated in the selected feature subset for several well-known clustering algorithms.

3.1 Parameter Setup

We set the parameters for our method. Assume that we have a dataset including a total of N instances. We first set an initial learning rate $\alpha(1)$ to 0.8 and decrease the learning rate $\alpha(t) = \alpha(1) \times [(T\text{-}t) / T]$ at the t^{th} iteration. The weight of every element in the initial feature salience vector $\mathbf{w}(1)$ is set to be the same. The number of iterations T should be large, such that most instances can be randomly selected for training on the algorithm. We thus set T to $10N$, where N is the size of the training dataset. The parameters K and L are set to $v = \{10, 20, \ldots, 100\}$ which depends on the training instances and will be discussed in later experiments.

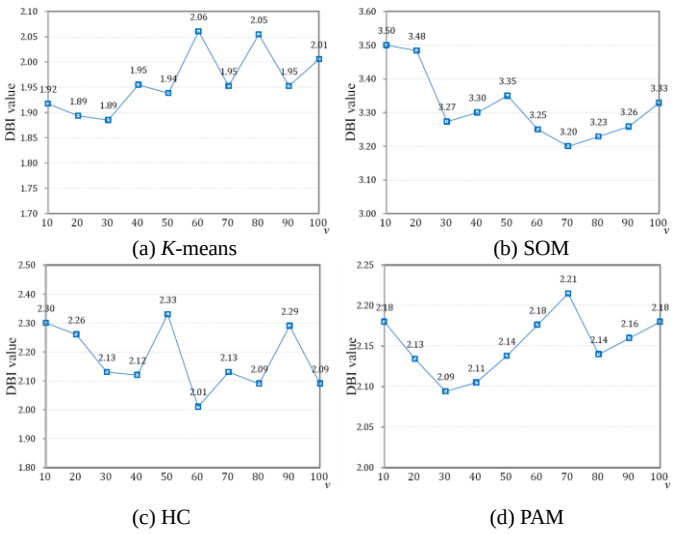
We used a dataset, OT, is mentioned in the studies [4, 15]. We followed the study [4] to randomly select 100 instances for each category and to capture features for every selected image.

3.2 Parameter Analysis

We observed how the parameter v affected the performance of our method because different size of v may produce differently salient features. Here, we discuss the performance of our presented methods for performing clustering (i.e., using K-means, SOM, HC and PAM) in the OT dataset. Each clustering algorithm partitioned the OT dataset into M clusters, where M was set according to the number of class labels (e.g.,

M = 8). With SOM, we followed the study [16] to obtain the user-defined number of clusters for the SOM units using K-means because similar units could be grouped. For the evaluation, we used the Davies-Bouldin index (DBI) [17] to measure the performance due to its popularity in cluster analysis. Lower DBI values represent higher compactness for the instances within a cluster, or higher separability for the instances between clusters. The comparison results of DBI values for various v for performing clustering in the OT dataset are shown in Figure 2.

In Figure 2, we can see that the size of v for performing K-means should be smaller so as to improve the performance. Specifically, the size of v for performing SOM and HC should be respectively set to 70 and 60, while the size of v in PAM should be 30. Therefore, we can set v to 30 rather than 40 for future analysis because v depends on the cost of searching the nearest and the farthest neighbors for every instance.



(a) K-means (b) SOM

(c) HC (d) PAM

Figure 2. Comparison of DBI values for various v in the OT dataset. The salient features were selected by our method and were used to perform clustering using (a) K-means, (b) SOM, (c) HC and (d) PAM algorithms. The parameter v = {10, 20, …, 100} was set to observe the corresponding performance.

IV. CONCLUSIONS

We present a new feature selection method that uses instance-based learning for quantifying features in the context of the nearest and the farthest neighbors of every instance. Such a method is advantageous because our method does not need to explore natural clusters using a predetermined clustering algorithm. Our comprehensive experiment on a synthetic dataset demonstrates that the salient features can be extracted effectively. Because the cluster analysis for very high data dimensionality has been in high demand, we expect that our work will generate broad interest in many research fields.

ACKNOWLEDGMENTS

The author would like to thank the reviewers for their valuable suggestions. This research was supported by National Science Council, Taiwan under grant NSC 99-2410-H-146-001-MY2.

REFERENCE

[1] Manning, C. and Schutze, H., *Foundations of statistical natural language processing*: MIT Press, 1999.

[2] Swets, D. L. and Weng, J. J., "Efficient content-based image retrieval using automatic feature selection," in *Proceedings of 1995 IEEE International Symposium on Computer Vision* 1995, pp. 85-90.

[3] Law, M. H. C., Figueiredo, M. A. T., and Jain, A. K., "Simultaneous feature selection and clustering using mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26(9), pp. 1154-1166, 2004.

[4] Boutemedjet, S., Bouguila, N., and Ziou, D., "A hybrid feature extraction selection approach for high-dimensional non-Gaussian data clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 3(8), pp. 1429-1443, 2009.

[5] Sanguinetti, G., "Dimensionality reduction of clustered data sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30(3), pp. 535-540, 2008.

[6] Dy, J. G. and Brodley, C. E., "Feature selection for unsupervised learning," *Journal of Machine Learning Research*, vol. 5, pp. 845-889, 2004.

[7] Chow, T. W. S. and Huang, D., "Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information," *IEEE Transactions on Neural Networks*, vol. 16(1), 2005.

[8] Peng, H., Long, F., and Ding, C., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(8), pp. 1226-1238, 2005.

[9] Kwak, N. and Choi, C. H., "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13(1), pp. 143-159, 2002.

[10] Pena, J. M. and Nilsson, R., "On the complexity of discrete feature selection for optimal classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32(8), pp. 1517-1522, 2010.

[11] Pal, M. and Foody, G. M., "Feature selection for classification of hyperspectral data by SVM," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48(5), pp. 2297-2307, 2010.

[12] Yang, J. B., Shen, K. Q., Ong, C. J., and Li, X. P., "Feature selection for MLP neural network: The use of random permutation of probabilistic outputs," *IEEE Transactions on Neural Networks*, vol. 20(12), pp. 1911-1922, 2009.

[13] Haykin, S. S. and Widrow, B., *Least-mean-square adaptive filters*: Wiley, 2003.

[14] Macchi, O., *Adaptive processing: The LMS approach with applications in transmission*: New York: Wiley, 1995.

[15] Oliva, A. and Torralba, A., "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145-175, 2001.

[16] Vesanto, J. and Alhoniemi, E., "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11(3), pp. 586-600, 2000.

[17] Davies, D. L. and Bouldin, D. W., "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1(4), pp. 224–227, 1979.

 ISBN: 978-1-61208-217-2