EXTRACTING THE SIGNIFICANT DEGREES OF ATTRIBUTES IN UNLABELED DATA USING UNSUPERVISED MACHINE LEARNING

Byoung Jik Lee

School of Computer Sciences, Western Illinois University
Macomb, IL, U.S.A.

ABSTRACT

We propose a valid approach to find the degree of important attributes in unlabeled dataset to improve the clustering performance. The significant degrees of attributes are extracted through the training of unsupervised simple competitive learning with the raw unlabeled data. These significant degrees are applied to the original dataset and generate the weighted dataset reflected by the degrees of influentialvalues for the set ofattributes. This work is simulated on the UCI Machine Learning repository dataset. The Scikit-learn K-Means clustering with raw data, scaled data, and the weighted data are tested. The result shows that the proposed approach improves the performance.

KEYWORDS

Unsupervised MachineLearning, Simple Competitive Learning, SignificantDegree of Attributes, Scikit-learn K-Means Clustering, Weighted Data, UCI Machine Learning Data.

1. INTRODUCTION

Data is extremely valuable to our lives. Data are being collected and saved almost anywhere and anytime. The size of data is increasing exponentially over time. Especially, the data that comes without a label, unlabeled data, is growing in greater volume and at a faster rate than labeled data. The unlabeled data is essential for unsupervised machine learning. Unsupervised learning with unlabeled dataset has a limited scale of computation and performance compared to supervised learning with labelled dataset. However, clustering, a process of partitioning a given data set into distinct groups, has been applied to the varieties of data mining, knowledge discovery, and pattern recognition applications [1] [2].

To enable data to be used for machine learning, some steps of data preprocessing are required such as data cleaning, data integration, data reduction, and data transformation [3].Normalization is a one of the most popular methods in data preprocessing. Attributes values are typically normalized by scaling original data within the specified range of values. When the range of the attribute values are significantly varied, normalization process has been used to balance the importance of the attributes of the data.

However, if some attributes have significant impacts on data clustering, it is advisable to treat these important attributes as meaningful core properties for clustering. There are two approaches to deal with the set of attributes. One method is to construct new attributes by adding new attributes or replacing current attributes from existing attributes. The other method is to deduct

David C. Wyld et al. (Eds): COMIT, SIPO, AISCA, MLIOB, BDHI - 2020
pp. 93-98, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101608

the irrelevant or redundant attributes [4][5][6]. Both methods change the set of existing attributes by increasing or decreasing the set of attributes. This can lead to additional inconsistency issues when new data are added in the future.

In this paper, to improve the clustering performance, we propose a method of assigning significant degree to important attribute sets while maintaining the set of attributes. In section 2, unsupervised simple competitive learning [7][8] is used to extract the estimated significant degree for each attribute, S_j, from the result of training with unlabeled data. This significant degree, S_j, is applied to the original dataset and produces the weighted dataset.

To verify this proposed method, K-means clustering, one of the most widely used method because of its simplicity to use and its high efficiency of computation, is employed. K-means clustering has been successfully applied to the variety of applications [9]. In section 3, the K-Means imports the trained weighted dataset and tests the popular data set. The result of Iris, Seeds, and Wine dataset of UCI Machine Learning Repository [10] shows that the proposed approach improves the performance.

2. SYSTEM ARCHITECTURE

This system has two modules. In Module 1, unsupervised simple competitive learning trains the raw unlabeled data and produces the estimated significant degree for each attribute, S_j. In Module 2, the K-Means clustering with raw data, scaled data, and the weighted data by the significant degrees are tested.

2.1. Module 1: Finding out the significant attributes by Simple Competitive Learning

Figure 1 shows the architecture of Simple Competitive Learning. The unlabeled data with *j* number of attributes are fed into the same number of input units in the network. The number of output units, *i*, is the number of clustering groups. The winner is the output unit with the largest net input for the feeding input vector *X*. The solid line represents an excitatory connection and the dashed line which connects output unit each other represents an inhibitory connection.

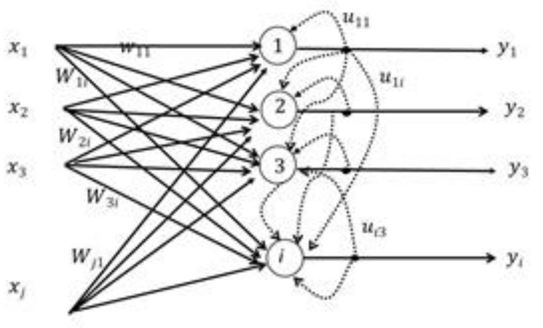

Figure 1. Simple Competitive Learning Network

With the given unlabeled dataset, the Simple Competitive Learning model is trained based on the learning algorithm (1) followed by the normalization process to avoid no bounds on the weight.

$$W_{ji}(t+1) = W_{ji}(t) + n\,(X_j^u - W_{ji}(t))\,, \qquad (1)$$

Where n is the learning rate, u is the u_{th} training data, t is the time step.When the training is completed, the weight W_{ji} represents the strength weight between the attribute j and the clustering group i. The significant degree of each attribute $j$is computed from the equation (2).

$$S_j = \sum_I W_{ji} \qquad (2)$$

The significant vector S constitutes the significant degree of each attribute S_j, $S = (S_1, S_2, ..., S_j)$.The significant degree of each attribute, S_j, is applied to the unlabeled dataset, X_j^u, and produces the weighted data, $Weighted\_data_j^u$, which embedded the significant degree of attribute j.

$$Weighted\_data_j^u = X_j^u \times S_{j,} \qquad (3)$$

Where j is the corresponding attribute, u is the u_{th} training data.

2.2. Module 2: K-Means clustering with raw data, scaled data, and weighted data

We used the Scikit-learn library [11] to explore K-Means clustering performance with the raw data, scaled data, and the proposed weighted data. Figure 2 illustrates the procedure of the system. The popular Elbow method of Scikit-learn library [11] was used to decide the number of clusters of the Simple Competitive learning Module 1.

```
Input: Unlabeled dataset
Output: Significant degree of each attribute$j$, $S_j$
    0. Decide the number of clusters by Elbow method
    1. Feed the unlabeled data into Simple Competitive Learning Network
    2. Train the Simple Competitive Learning network based on the learning equation (1)
    3. Compute the Significant degree of each attributes, $S_j$, by equation (2)
    4. Apply the Significant degree to the unlabeled data to produce the Weighted data by
       equation (3).
    5. Explore K-Means clustering with raw data, scaled data, and the proposed weighted data.
```

Figure 2. Exploring the Significant Degree of Attributes Procedure

3. EXPERIMENT RESULT

3.1. The Dataset and the Significant Degree of the Data

UCI Machine Learning repository dataset was used for performance evaluation. Wine dataset, Iris dataset, and Seed datasets are accessed to verify the effect of the proposed approach. The Iris data has 150 instances, four attributes (length and width of sepals and petals), and three classes. Each class (Iris setosa, Iris virginica, and Iris versicolor) has 50 instances. The input values of the attributes are in centimeters. Wine data has 13 attributes (Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines and Proline), 178 instances, and 3 classes. The input values of attributes are continuous. The Seed data has seven attributes (area A, perimeter P, compactness C, length of kernel, width of kernel, asymmetry coefficient, length of kernel groove), 210 instances, and 3 classes. The input values of all attributes are real-valued continuous.

As shown in Figure 3, three attributes (compactness C, asymmetry coefficient, length of kernel groove) of See data are the significant attributes for determining the clustering. The significant degree for Seed data is [0.3103793337657787, 0.38404893997441847, 0.5627959563346043, 0.38879262126954195, 0.35041866849291087, 0.5255126477711352, 0.47805183239161064].

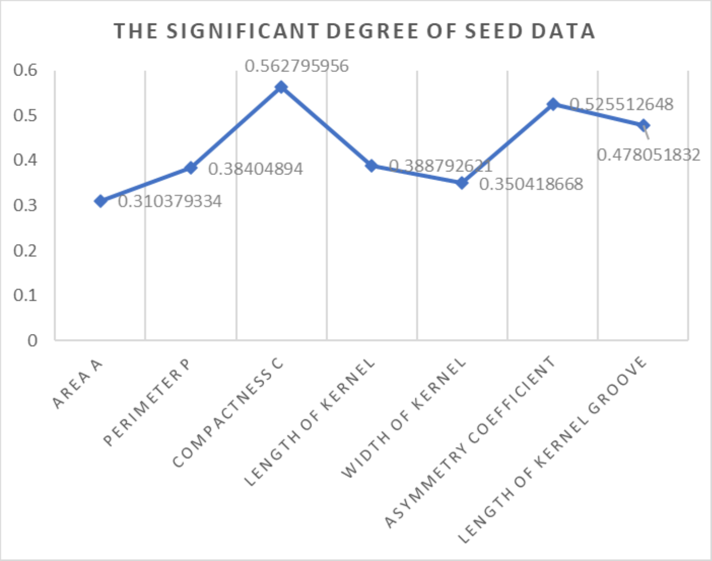

Figure 3. The significant degree of Seed data

As shown in Figure 4, the last attribute (petal width in cm) of Iris data influenced to decide the clustering.The significant degree for Iris data is [0.7569583177646524, 0.6202633420788992, 0.7613989085834454, 1.1575052380511397].

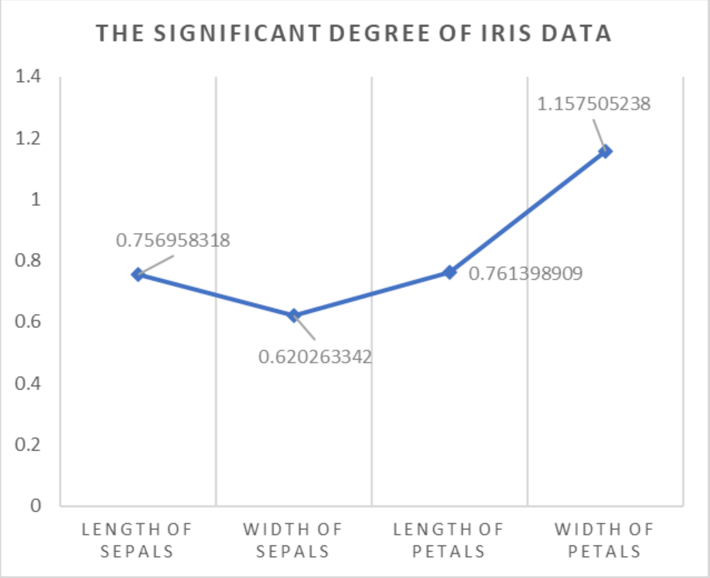

Figure 4. The significant degree of Iris data

The significant degree for Wine data is [0.49679553839142865, 0.4253067347939007, 0.3365439062186957, 0.2845606505403928, 0.1526467495303905, 0.2809781772430352, 0.2718178595519532, 0.20136617303733317, 0.27474333797269745, 0.13341331581839957, 0.25327013905125795, 0.37015396979853094, 0.20009014838009148]. As shown in Figure 5, the first attribute (Alcohol) and the second attribute (Malic acid) of Wine data significantly contributes to determining three types of wine.

3.2. Improvements and Limitations

Table 1 and Table 2 show that the proposed approach improves the performance of clustering problem in three datasets. Comparing the scaled data with the suggested weight data, the performance of the Iris and Seed data are improved by 4.7% and 3.4%, respectively. Of the three problems (Wine, Iris, Seed), the proposed approach is the most effective in the Iris problem, because one attribute (petal width in cm) has a high significant degree than other attributes. As observed in 1.7% improvement of Wine data, this approach has limitations if there are not significant attributes in the data.

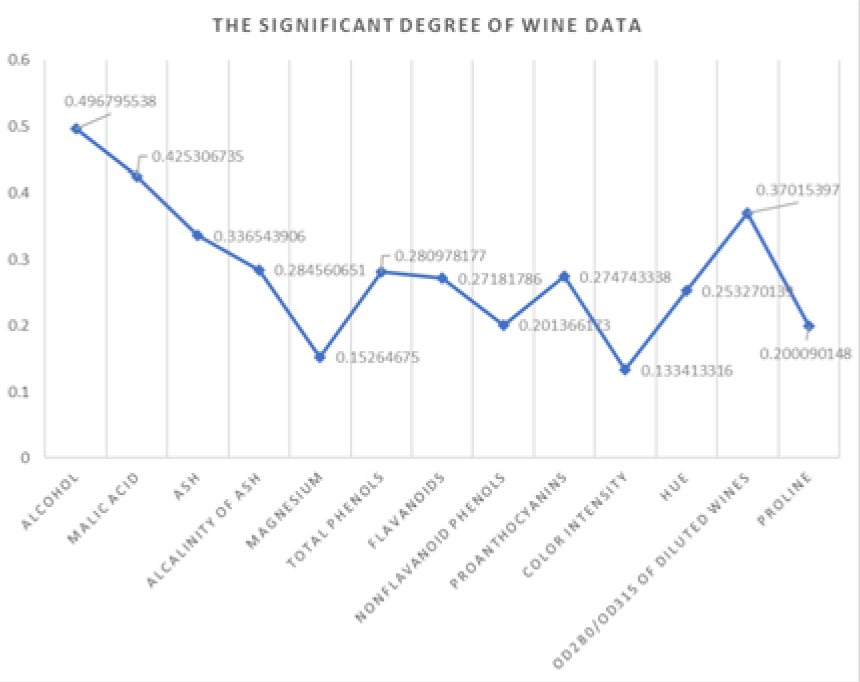

Figure 5. The significant degree of Wine data

Table 1. Performance of raw data, scaled data, weighted data

Models	Correction rate		
	Wine	*Iris*	*Seeds*
Raw data	71.3%	89.3%	89.5%
Scaled data	95.5%	89.3%	89%
Weighted data	97.2%	94.0%	92.4%
Scaled data to Weighted data	1.7%	4.7%	3.4%

Table 2. Performance of raw data, scaled data, weighted data

		Wine			Seeds			Iris		
		A	B	C	A	B	C	A	B	C
Raw data	A	46	0	13	60	1	9	50	0	0
	B	1	51	19	10	60	0	0	48	2
	C	0	18	30	2	0	68	0	14	36
Scaled data	A	59	0	0	58	2	10	50	0	0
	B	2	63	6	8	68	0	0	48	2
	C	0	0	48	3	0	67	0	14	36
Weighted data	A	58	0	1	64	2	4	50	0	0
	B	2	67	2	3	67	0	0	45	5
	C	0	1	47	7	0	63	0	4	46

4. CONCLUSIONS

This paper proposes an approach to extract the significant degree of each attribute in unlabeled dataset for efficient clustering performance while maintaining the set of attributes.This significant degree of each attribute converts the unlabeled raw data into the weighted data which reflects the importance of each attribute. This approach is tested by Scikit-learn K-Means clustering on some of UCI Machine Learning repository dataset with raw data, scaled data, and weighted data. The result shows that the proposed approach improves the performance for all tested data sets.

REFERENCES

[1] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in KnowledgeDiscovery and Data Mining. AAAI/MIT Press, 1996

[2] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973

[3] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012

[4] H. Liu and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic, 1998.

[5] W. Siedlecki and J. Sklansky. On automatic feature selection. Int. J. Pattern Recognition and Artificial Intelligence, 2:197–220, 1988.

[6] Pyle, D., 1999. *Data Preparation for Data Mining.* Morgan Kaufmann Publishers, Los Altos, California.

[7] J. Hertz, A. Krogh, R. Palme, Introduction to the Theory of Neural Computation, Addison Wesley, 1991

[8] D. E. Rumelhart and J. L. McClelland, Parallel Distributed Processing, MIT Press, 1986,, pp. 151-193

[9] M M. N. Murty, A. K. Jain and P. J. Flynn, "Data Clustering: A Review", ACM Computing Survey, Vol. 31, No. 3, 1999, pp 264-323.

[10] D. J. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine, CA, 1988

[11] S. Rashika and V. Mirjalili, Python Machine Learning. 2nd Edition. Birmingham, UK: Packt Publishing, 2017