Noname manuscript No.
(will be inserted by the editor)

An introduction to a systematic approach for producing meaningful clustering results.

Navid Yousefabadi

April 20, 2020

Abstract Unsupervised learning algorithms are used to divide a dataset into clusters. The aim of the clustering is to produce meaningful clusters. However, the result of the clustering is usually not consistent. The same dataset could be divided into meaningful groups and labeled by a domain expert. In this work we show preliminary results for a systematic approach to produce meaningful clustering results without having access to a domain expert opinion. This was done by testing a hypothesis, which states that increasing the relevant feature space dimensions, results into output clusters that more accurately and precisely match with a domain expert labeling. This is conclusive evidence that with increased dimensionality of relevant features, meaningfulness emerges as a property of a clustering system.

Keywords Unsupervised Leaning · Meaningful Clustering · K-means · Hierarchical Clustering · HDBSCAN · GMM · Hypothesis Test

1 Introduction

Unsupervised learning is a type of machine learning that has important applications in pattern recognition and data mining with no pre-existing labels and minimum human supervision. Unsupervised learning algorithms have been developed in various fields. In the area of image processing it has been used to build a face detector without having to label images as containing a face or not [1]. In the field of bioinformatics broad patterns of gene expression is revealed by clustering analysis [2], [3]. Or in the field of natural language processing it has been used for text summarization [4]. In all of these works a specialized unsupervised learning algorithm is developed and tested to produce meaningful results. However, there is no systematic way to ensure meaningfulness of the results without assessing the clustering outcomes. Although unsupervised learning algorithms were developed with the promise of minimum human supervision. The results could be inconsistent, and finding a meaningful clustering might not be trivial. There is always a need to inspect the results in order to make interpretations. A domain expert might be needed to interpret the results, if a meaningful clustering was produced in the first place. We state that one way to produce meaningful clusters is to use domain expert labeling, since annotations of a domain expert are always meaningful. We then seek to find conditions where the results of clustering become similar to a domain expert labeling. In order to provide a systematic approach, we investigate unsupervised learning holistically as a system. The components of an supervised learning system could be a collection of datasets, clustering algorithms, hyperparameters and etc. This will generalize our results for any unsupervised learning system.

A hypothesis testing was performed to provide evidence that in a system of unsupervised learning increasing the number of relevant features always leads to results that more accurately and precisely match with domain expert labeling. A conclusive evidence that increasing the number of relevant features always leads to production of meaningful clusters. It should be noted that our claim was only tested with the Iris dataset with four feature space dimensions and is only valid for this dataset. In order to generalize our results as an emergent property of unsupervised learning systems, this hypothesis should be tested for any collection of dataset

N. Yousefabadi
University of Calgary
E-mail: navid.yousefabadi@ucalgary.ca

type and clustering algorithm and other components that are needed as a part of clustering algorithm.

In Sect. 2 an overview of the method is provided. In Sect. 2.1 datasets, and in Sect. 2.2 four different unsupervised learning algorithms that were used for this analysis are provided. In Sect. 4 the results of all of the algorithms that were produced with different numbers of feature space dimensions is compared to see if the number of feature space dimensions has a statistical significance in the meaningfulness of the outcomes.

2 Method

A systematic approach refers to an analysis method that handles a complex system with a global point of view without focusing on the details. Unlike analytical approach where factors responsible for a property are identified, in a systematic approach emerging properties of a system are identified. This approach is used in a variety of domains such as information systems, decision systems, biological systems, etc. For example a systems approach is required in order to identify a sustainable agriculture [14]. In a systems approach, knowledge-based development of whole farms and communities will be required to address the environmental, economic, and social challenges of the post-industrial era of agricultural sustainability [14].

In order to provide a systematic approach for producing meaningful clusters, unsupervised learning is considered as a whole system. Some of the constituents of a clustering system could be listed as the choice of unsupervised learning algorithm, the assessment of the dataset, hyperparameter tuning, interpretation of the results, etc. In a systematic approach, instead of focusing on individual algorithms and datasets as components of this system, it is investigated holistically.

By considering the clustering analysis as a whole system, the emerging property of meaningfulness is quantified as following. It can be stated that the expert labeling is at least one way to produce meaningful results. The measure of meaningfulness can be defined as the similarity measure between clustering results and the expert labeling. The matching accuracy was chosen as a measure of similarity between domain expert labeling and the clustering results. The matching accuracy is calculated by dividing the number of clustering instances that match the domain expert labeling over the total number of instances. The matching precision is the range of the accuracy. The goal is to find conditions that always lead to producing results that match with a presupposed domain expert labeling. In order to provide statistical evidence for meaningfulness, an emergent property of the unsupervised learning systems, a variety of different algorithms and datasets were used. We show that in spite of having different algorithms and datasets, the matching accuracy and precision increases with the number of relevant attributes.

One way a domain expert would label the data would be to look at the different columns of the data and set ranges of columns (for numerical columns) or pick values of columns (for categorical columns) that correspond to specific scenarios or objects. These scenarios or objects form meaningful clusters of data. Different columns are the attributes of objects or scenario. One can make an educated guess and say that the separation between objects and scenarios becomes more obvious by increasing the number of relevant attributes that pertain to those objects or scenarios.

A hypothesis could be formed to provide evidence that by increasing the dimensionality of the feature space the matching accuracy and precision with domain expert labeling increases. The hypothesis could be formulated as following:

$$H_0 : \bar{X}_i \leq \bar{X}_{i+1} \qquad\qquad H_A : \bar{X}_i > \bar{X}_{i+1} \tag{1}$$

Where $\bar{X}_i$ is the mean matching accuracy using i relevant feature space dimensions. The null hypothesis H_0 states that the mean matching accuracy using i relevant feature dimensions is less or equal than the mean matching accuracy using $i + 1$ relevant feature dimensions. The alternative hypothesis H_A states that the mean matching accuracy using i relevant feature dimensions is greater than the mean matching accuracy using $i+1$ relevant feature dimensions. Permutation testing,a non-parametric hypothesis test, was used used. Since the difference between means of the two populations does not have a normal distribution and permutation is a condition free hypothesis test.

2.1 Datasets

In order to generalize our finding a number of various dataset types shall be used. If the matching precision increases with feature dimensions for all data types, it could be concluded that our finding is independent of dataset. Popular datasets available on R packages were chosen. Iris dataset that gives the sepal length and width and petal length and width of 3 spices of flowers: Setosa; Virginica; and Versicolor Fig. 1. The Sonar data from mlbench package , where rock and mines are identified with 60 features. The features of Sonar data represent energy within particular frequency bands. The costumer churn data from C50 package with 19 features with which churn and no churn costumers can be identified.

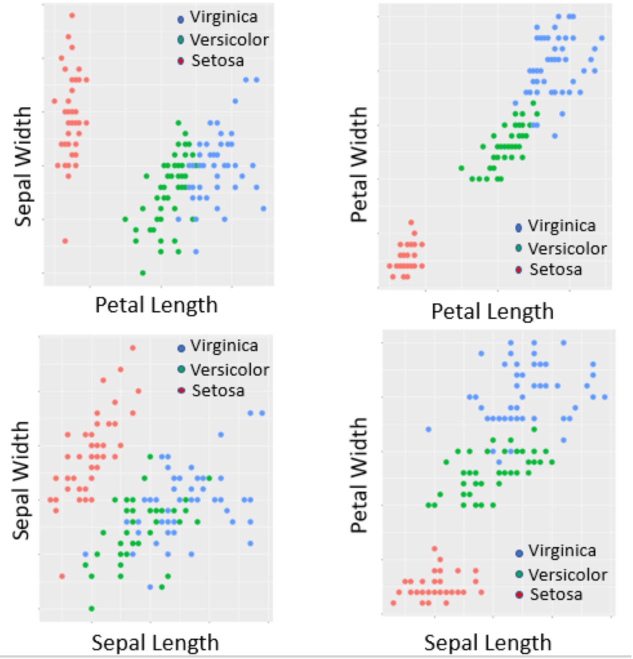

Fig. 1: Two dimensional plot of the iris dataset. The abscissa and the ordinate showing petal and sepal width and length. The color shows the spices of the flower. The labeling of the spices was provided by a domain expert.

2.2 Unsupervised learning algorithms

In order to generalize our finding a number of various unsupervised learning algorithms shall be used. If the matching precision increases with feature dimensions for all learning algorithms, it could be concluded that our finding is independent of the choice of learning algorithm. Popular unsupervised learning algorithms such as k-means, hierarchical clustering, hierarchical density-based spatial clustering of applications with noise (HDBSCAN), and Gaussian mixture model (GMM) that are available on R packages were used.

The hyperparametrs of the algorithms were tuned appropriately as a prerequisite for producing meaningful clusters. Some of the important questions when assessing the suitability of an unsupervised learning algorithm for a dataset are as follows. How is the separation between possible clusters? Linear, hyperspherical, etc. How many clusters could be expected from the data? What is the minimum number of instances per cluster? Could instances belong to more than one cluster or not? (soft clustering or hard clustering) [5]. In order to not lose generality we used variety of algorithms suitable for various types of data. The number of clusters and the minimum number of instances per cluster was determined as was required for algorithms. Algorithms suitable for both hard and soft clustering were used. The aim was to generalize our results for any unsupervised learning algorithm and data type.

K-means is one of the most popular clustering algorithms [7]. This algorithm is centroid based and suitable for clustering data with hyper spherical shape. It is a hard clustering algorithm for which the number of clusters should be determined. K-means has time complexity of O(n) and can handle big data well [5]. Hierarchical clustering is a method of clustering in which a hierarchy of clusters is built [6] and the results are usually presented in a dendrogram. The number of clusters can be set by making a horizontal cut in the resulting dendrogram. HDBSCAN, a density based clustering algorithm was first proposed by Martin Ester et.al in 1996 [8]. DBSCAN is one of the most common and most cited clustering algorithms in scientific literature [9]. The algorithm was awarded the test of the time award in 2014 [10]. The only hyperparameter that should be determined for this algorithm is the minimum number of instances in one cluster. This algorithm is a soft clustering and can fit to non-linearly separable data. GMM a variant of mixture models uses a weighted sum of multivariate Gaussian distributions to model the data [5]. This algorithm is used in a vast number of applications. In finance it is used to model financial return in normal situations and crisis [11]. In image processing it is used for handwriting recognition [12]. In predictive maintenance and early fault detection [13] mixture model-base clustering is predominantly used for identifying the state of the machine e.g normal state, power off state, or faulty state. GMM is also a soft clustering algorithm where the clusters have the form of an ellipse that can have arbitrary elongation and rotation [5]. The number of clusters is also a hyperparameter that should be determined in GMM.

3 Results

Our results were produced using only the Iris dataset. However, a conclusive statistical evidence requires this analysis to be performed on a number of datasets which should exceed 34. In order to provide evidence for our claim we need to cluster the data using different feature space dimensions. The Iris dataset has 4 dimensions. The total number of possibilities for choosing r dimensions between n dimensions is $\frac{n!}{n!(n-r)!}$.

The accuracy for matching with the domain expert labeling can be plotted, after performing the clustering for all different possible combinations of the attributes, and clustering algorithms. Fig. 2 summarizes the results for matching accuracy vs feature space dimensions using all possible feature space dimensional possibilities, and the unsupervised learning algorithms mentioned in section Sec. 2.2.

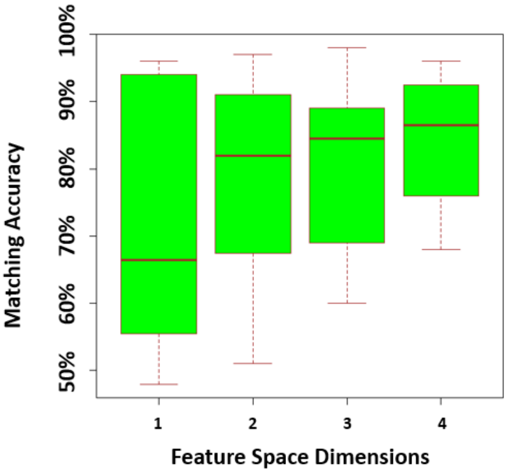

Fig. 2: Matching accuracy when matching the clustering results with domain expert labels

As it can be seen in the Fig. 2 box-plot of accuracy for different combinations of feature space dimensions, the accuracy increases by increasing the feature space dimensions. The standard deviation of the accuracy also decreased by increasing the dimensions which shows that the precision is increased.

The permutation test was performed by sampling from the results of clustering matching accuracy 30000 times. In Fig. 3 three plots can be seen which are the results of three hypothesis tests. A permutation test between two and three relevant feature space dimensions, a test between two and three dimensions, and a test between three and four dimensions. The p-values for each test can be seen in the figure. In all three tests there is not enough evidence to reject the null hypothesis. These results are evidence that by increasing the number of relevant feature space dimensions the matching accuracy also increases. If no evidence can be found to reject our hypothesis, it can be concluded that by increasing the number of relevant feature space dimensions, the mean matching accuracy with the domain expert labeling is going to increase. Therefore, by having enough relevant feature space dimensions in an unsupervised learning system the clustering results are going to accurately match with a domain expert labeling, although the domain expert labeling might not be available.

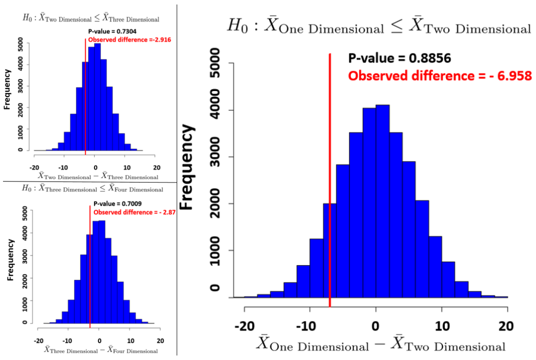

Fig. 3: Permutation results for three hypothesis tests: tests between one and two relevant feature space dimensions, two and three dimensions and three and four dimensions. The red vertical line indicates the observed difference. The p-value for each test is reported on top right corner of each plot.

If the uncertainty of the mean matching accuracy was large, the result of the hypothesis testing could be different. The plot of the matching precision versus relevant feature space dimensions can be seen in Fig. 4. An increasing trend for matching precision could also be observed. It would be interesting to test the hypothesis that by increasing feature space dimensions the matching precision also increases. However, this testing was not performed.

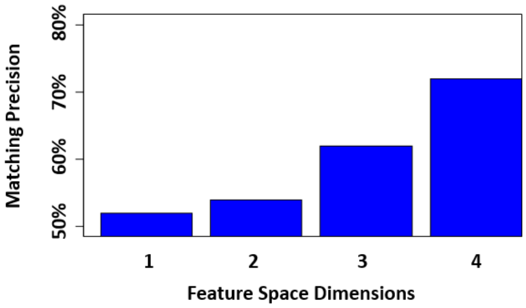

Fig. 4: Matching precision versus relavent feature space dimensions.

4 Conclusion

It can be concluded that by having enough relevant feature space dimensions, regardless of the unsupervised learning approach, the clustering results will accurately match with a domain expert labeling. We provided statistical evidence that even if there is no expert labeling available, it can be guaranteed that by having enough relevant feature space dimensions the clustering results are going to resemble that of a domain expert labeling and have a property of meaningfulness.

5 Future Work

The statistical evidence provided in this work is limited to Iris dataset and its four dimensions. The unsupervised learning algorithms were also limited to four algorithms listed in Sec. 2.2. In order to not lose generality for unsupervised learning as a system a variety of other datasets and clustering algorithms should be included in the analysis.

A very interesting question that could be pursued is that how many relevant feature space dimensions is enough in order to produce results that would resemble that of a domain expert? There has been a number of works done on feature selection for unsupervised learning [15], [17], [16]. Our result has made it evident that choosing enough relevant features leads to clustering results that resemble that of a domain expert. It would be interesting to provide statistical evidence for the hypothesis that unsupervised learning systems with proper feature selection component are going to produce meaningful results.

References

1. Q. V. Le,Building high-level features using large scale unsupervised learning, 8595-8598. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (2013)
2. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Science, 96(12):6745–6750, June 1999
3. A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. Journal Of Computational Biology. 6:281-297, 1999.
4. R. A. García-Hernández, R. Montiel, Y. Ledeneva, E. Rendón, A. Gelbukh, and R. Cruz, Text Summarization by Sentence Extraction Using Unsupervised Learning, MICAI 2008: Advances in Artificial Intelligence, Springer Berlin Heidelberg, 133–143, 2008.
5. Andriy Burkov, The Hundred-Page Machine Learning Book, 30-42. Kindle Direct Publishing, 2019.
6. Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. 321-352 Springer US, 2005.
7. Lloyd, S. Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129–137, 1982. doi:10.1109/tit.1982.1056489
8. M. Ester, H. P. Kriegel, J. Sander, and X. Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, KDD 1996.
9. "https://web.archive.org/web/20100421170848/http://academic.research.microsoft.com/CSDirectory/paper_category_7.htm". Retrieved 2020-04-20. Most cited data mining articles according to Microsoft academic search; DBSCAN is on rank 24.
10. "https://www.kdd.org/News/view/2014-sigkdd-test-of-time-award". Retrieved 2020-04-20. ACM SIGKDD. 2014-08-18.
11. Dinov, ID. Expectation Maximization and Mixture Modeling Tutorial. California Digital Library, Statistics Online Computational Resource, Paper EM_MM, http://repositories.cdlib.org/socr/EM_MM, 2008
12. C. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag New York 2006
13. Amruthnath, Nagdev; Gupta. A Research Study on Unsupervised Machine Learning Algorithms for Fault Detection in Predictive Maintenance. Unpublished 2018. doi:10.13140/rg.2.2.28822.24648
14. J. Ikerd, The need for a system approach to sustainable agriculture. Agric Ecosyst Environ 46(1-4):147-160. Agriculture, Ecosystems & Environment. 46. 147-160. 10.1016/0167-8809(93)90020-P. 1993.
15. J.Yao, Q. Mao, S. Goodison, V. Mai, Y. Sun, Feature selection for unsupervised learning through local learning, Pattern Recognition Letters, Volume 53, 2015, Pages 100-107, ISSN 0167-8655, https://doi.org/10.1016/j.patrec2014.11.006.
16. J. G. Dy, C. E. Brodley; Feature Selection for Unsupervised Learning, 5(Aug):845–889, 2004.
17. J. Handl, J. Knowle, Feature Subset Selection in Unsupervised Learning via Multiobjective Optimization, International Journal of Computational Intelligence Research, ISSN 0973-1873 Vol.2, No.3 (2006), pp. 217–238