Cross-Validation Approach to Evaluate Clustering Algorithms: An Experimental Study Using Multi-Label Datasets

Cross-Validation Approach to Evaluate Clustering Algorithms: An Experimental Study using Multi-label Datasets

Adane Nega Tarekegn [1], Krzysztof Michalak [2], Mario Giacobini [3]

[1]Modelling and Data Science, Department of Mathematics, University of Turin, Italy
Email: adanenega.tarekegn@unito.it

[2]Department of Information Technologies, Wroclaw University of Economics, Poland
Email: krzysztof.michalak@ue.wroc.pl

[3]Data Analysis and Modeling, Department of Veterinary Sciences, University of Turin, Italy,
Email: mario.giacobini@unito.it

Abstract

Clustering validation is one of the most important and challenging parts of clustering analysis, as there is no ground truth knowledge to compare the results with. Up till now, the evaluation methods for clustering algorithms have been used for determining the optimal number of clusters in the data, assessing the quality of clustering results through various validity criteria, comparison of results with other clustering schemes, etc. It is also often practically important to build a model on a large amount of training data and then apply the model repeatedly to smaller amounts of new data. This is similar to assigning new data points to existing clusters that are constructed on the training set. However, very little practical guidance is available to measure the prediction strength of the constructed model to predict cluster labels for new samples. In this study, we proposed an extension of the cross-validation procedure to evaluate the quality of the clustering model in predicting cluster membership for new data points. The performance score was measured in terms of the root mean squared error based on the information from multiple labels of the training and testing samples. The principal component analysis (PCA) followed by k-means clustering algorithm was used to evaluate the proposed method. The clustering model was tested using three benchmark multi-label datasets and has shown promising results with an overall RMSE of less than 0.075 and MAPE of less than 12.5% in three datasets.

Keywords: Clustering validation, Clustering analysis, Cross-validation, Multi-label data

1. Introduction

1.1. Overview of Unsupervised Learning

Unsupervised learning aims to find the underlying structure or the distribution of data. It is an important area in the domain of machine learning, where the labels for the data examples are not necessarily required for model building. The main tasks in unsupervised learning include cluster analysis [1,2], building self-organizing maps (SOM)[3], representation learning [4], and density estimation [5]. Cluster analysis, the main focus of this study, is a central task for grouping heterogeneous data points into a number of more homogenous subgroups based on distance, or naturally occurring trends, patterns, and relationships in the data. The formation of homogenous or heterogeneous grouping (or clustering) structure from a complex dataset requires a measure of 'closeness' or 'similarity.' In clustering, the definition of similarity is highly dependent on the applied distance function between the data objects. The choice of similarity measure can be considered based on the type of the variable used to cluster objects (continuous, discrete, binary), the type of measurements (nominal, ordinal, ratio, interval), and subject matter knowledge. The most commonly used distance measure in most clustering algorithms is the Euclidian distance [6]. Other measures include Minkowski's distance [7],Cosine distance [8],S-distance[9] ,etc.

The clustering problem has a clear goal of finding distinct groups or 'clusters' within the dataset. However, the notion of a 'cluster' has not been precisely defined, which has driven to the development of several clustering paradigms and several clustering algorithms within each paradigm [10]. The existence of different types of clustering algorithms poses difficulties in selecting the best algorithm for a particular task. Independent of the type of algorithm used, Kleinberg [11] proposes three properties that an ideal clustering algorithm should have so that it can be considered good: scale invariance, consistency, and richness. Scale invariance indicates that the clustering algorithm does not change its results when all distances between points are scaled by a constant factor. A clustering process is considered to be consistent when the clustering results do not change if the distances within clusters decreases and/or the distance between clusters increase. The richness criteria mean that the clustering function must be flexible enough to potentially produce any arbitrary partitions of the input dataset. According to Kleinberg's impossibility theorem [11], no clustering algorithm satisfies all three requirements simultaneously. This implies that it has been very difficult to develop a unified framework for validation of clustering methods and to reason about it at a technical level.

1.2. Multi-label Data

Several types of research in machine learning deal with the analysis of single-label data, where training instances are associated with a single label λ from a set of disjoint labels L. However, training samples in several application domains are often associated with a set of labels $Y \subseteq L$. Such datasets are called multi-label data. Multi-label datasets have been popular in various domains, such as protein function classification, medical diagnosis, emotion recognition, text classification, etc. For instance, a medical patient may be affected by more than one chronic disease: diabetes, hypertension, and fatty liver. We can cluster the patients into distinct groups, each with specific characteristics, and then the burden of these unwanted outcomes (diabetes, hypertension, fatty liver, etc.) can be identified to provide tailored interventions in each cluster. One of the common trends for solving supervised learning through the use of multi-label data is decomposing the multi-label problem into binary classification problems [12-14]. In unsupervised learning, we can use the labels information of the multi-label data for evaluation of the clustering algorithm. In this study, we used features for forming clusters and class labels for performance evaluation.

1.3. Cluster Validation

Cluster validation is one of the most important and challenging parts of cluster analysis, which involves the objective and quantitative assessment of clustering results [2]. One of the problems in cluster validation is that there is no clear notion as to what exactly the 'prediction error' is. Because of that, clusters are sometimes validated by ad hoc methods based on the application area. Due to the absence of the ground truth and the nature of the problem, cluster validation has not been well developed [15]. As a result, evaluating the performance of a clustering algorithm is not an easy task. Commonly, the evaluation process depends on the algorithm used to obtain clustering results, which resulted in the development of multiple evaluation techniques. Various methods have been suggested in the literature for cluster validation, including external validation, internal validation, relative criteria, and stability based approaches.

External Clustering Validity Methods: external validation index uses prior knowledge, such as externally provided class labels, to evaluate results of cluster analysis. External clustering validity approaches, such as Rand Index [16] and normalized mutual information [17], are used to measure the quality of clustering results by comparing the generated cluster labels with the pre-existing clustering (reference labels) structure, i.e., ground truth solution. If the result is in some way similar to the reference, the final output is regarded as a "good" clustering. The

external validation is straightforward when the closeness between two clusterings is well-defined. However, it has a basic caveat that the reference result is not given in most real-world applications. Therefore, external evaluation is generally used for synthetic data and for tuning clustering algorithms [18].

Internal Cluster Validity Methods: these are used to assess the goodness of the clustering structure without reference to the external information, using only the data themselves. Internal clustering validity methods measure the quality of clustering based solely on information intrinsic to the data; as a result, they have great practical application and numerous criteria have been proposed in the literature, such as Silhouette analyses [19], Calinski–Harabasz index [20], Davies–Bouldin [21]. The internal criteria are the most commonly used evaluation methods designed to compute the ratio of within-cluster scattering (compactness) and to between-cluster separation. Measures that grouped under this category have been designed for the validation of convex-shaped clusters (such as globular clusters), and fail when applied to validate non-convex clusters [22].

The relative approach: is performed by comparing two sets of clusters (usually built with similar algorithms but with different parameter settings) to determine which one is better. It's generally used for determining the optimal number of clusters.

Clustering Stability Approach: clustering stability measure is a slightly different approach used to assess the similarity of clustering solutions obtained by applying the same clustering algorithm on multiple independent and identically distributed samples. The intuitive idea behind the stability approach is that if we repeatedly sample data points from the population and apply the candidate clustering algorithm, then a good algorithm should produce clusterings that do not vary much from one sample to another [23]. In other words, the algorithm is stable with respect to input randomization. There are several studies to validate clusters by stability criteria [24-26]. In general, the existing validation criteria are useful for such tasks as determining the correct number of clusters in the dataset, verifying whether the clusters obtained are meaningful or are just an artifact produced by the algorithms, justifying why we choose some algorithms instead of others or assessing the quality of clustering solutions. However, in the literature, there is still a lack of methods to measure the ability of the clustering algorithm to predict cluster memberships for new data points.

Generally, evaluating clustering results has been historically expressed as the most challenging topic [27]. In fact, Jain and Dubes [28], in their classic book on clustering, stated that:

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Despite achievements observed in this particular area over the past several years, it is highlighted that the above statement still remains true. This has motivated us and other many researchers in the area to study, develop, and propose methods to the validation of clustering results, as there is room for further investigation in the area.

1.4. The focus of this paper

The primary aim of this paper is to measure the performance of a clustering model to predict cluster labels for new data points, given that the model is already constructed from the training data. For example, we have three existing clusters, C1, C2, and C3, and a new data point D1. The clustering model should assign D1 to one of the clusters, say C2. In this case, we want to know 'how good is the model on new data?' i.e., to what extent the model has correctly assigned D1 into C2.

The cluster validation idea presented in this study is different from the existing methods in that it focuses on measuring the prediction strength of a clustering algorithm by using the cross-validation procedure. The k-fold cross-validation method is used for simulating the situation when we have built the clustering model on some previously available data, and then we want to assign new data points to the previously built clusters. The prediction strength concept presented here, similarly, as the stability of the clusters, can be used for assessing the performance of a clustering method. Clustering stability results are mostly obtained based on perturbations introduced to the input data, such as sub-sampling or the addition of noise. Unlike in the other studies, the prediction strength of an algorithm introduced here is measured by incorporating information from several labels of multi-label data. Namely, the probability of occurrence of the labels in the training and testing data is calculated for each cluster. If label probabilities in the training and testing data are similar, the clustering can be considered as a good one. Thus, this study assumes that the clusters are already formed from the training data, and the aim is to measure how well the clustering model predicts the corresponding cluster labels for the test data based on their membership on the clustering results obtained from the training data.

This approach is motivated by medical applications in which we would like to assess the probability of various health problems in different patient groups. For example, the labels for the chronic dataset are diabetes, hypertension, and fatty liver, as indicated in section 1.3. Once the clusters are formed, the probabilities of the occurrence of these labels, i.e., diabetes, hypertension, and fatty liver are estimated in each cluster and compared between the training set and the test set. The aim is to measure how well we can predict the probabilities of these three outcomes in new patients (i.e., in the test data) based on their membership in the training clusters. In this paper, the k-fold cross-validation procedure is used to simulate such a scenario.

The k-fold Cross-Validation (CV) is one of the most commonly used model evaluation procedures in supervised learning. Unfortunately, it is challenging to apply CV to unsupervised learning, for example, to clustering validation. In this study, the k-fold CV procedure is adapted, by using labels from a multi-label dataset, to be applicable to unsupervised learning (i.e., clustering) for evaluating the performance of clustering algorithms. Following the k-fold cross-validation approach, the input data is randomly divided into k parts, of which k-1 parts are used to construct the model, and the remaining part is used as an evaluation set. Then, the prediction strength is used as a statistic for clustering stability. Thus, here we propose the use of the k-fold cross-validation procedure for evaluating the prediction strength of the clustering model using the information acquired from multiple labels.

The contributions of this study are: (1) a new cluster validity index is proposed that uses the information from multiple labels to evaluate the quality of clustering algorithms; (2) the study validates the proposal through the cross-validation analysis of some challenging multi-label datasets; (3) the root mean squared error (RMSE), which is the most frequently used measure of the differences between values in regression problem, is exploited and adjusted to be used as a cluster validity index; (4) this study shows that the proposed method can be used to measure the ability of a clustering algorithm to predict the cluster membership for new data.

2. Proposed Method

Given a particular clustering result, one can predict cluster membership for new data based on a clustering model built on training data. This is not always easy for all types of clustering algorithms. For example, it is hard for density-based clustering algorithms (e.g., DBSCAN) to predict a cluster for the new data points, because the new data points may change the underlying clustering structure. For centroid-based cluster algorithms (e.g., k-means clustering), however, the prediction of a cluster for new data points is relatively easy since it only requires finding

the minimum distance of a new data from all cluster centers and then updating the cluster center of that cluster. Hence, k-means clustering is employed to test the proposed method in this paper. Recently, several techniques have been proposed to improve the standard k-means algorithm for high dimensional datasets, such as the Entropy Weighted Power k-Means [29], sparse k-means [30] and others [31]. The proposed k-Fold CV for unsupervised learning can also be applied to these modified versions of the k-means algorithms. In hierarchical clustering [32], assigning new objects to the existent clusters can be challenging since hierarchical clustering doesn't partition the input data, rather it connects some of the objects given during clustering to build a hierarchy of clusters described by a dendrogram. Hence, the proposed clustering metric can be hardly applicable to the hierarchical clustering techniques. Moreover, the computational complexity of hierarchical clustering can be a bottleneck to apply k-fold CV well, particularly on large datasets. However, it is an open question as to whether the proposed clustering metric can be extended to the improved versions of hierarchical clustering methods, such as clustering with optimal transport [33], and whether such extensions will perform well in practice.

Assigning new data points to existing clusters that are constructed through the training data is considered to be an important practical application. However, very little practical guidance is available to measure the prediction strength of the constructed model to predict the cluster membership of a new data point. Prediction strength is a global measure forcing all clusters to be stable, as it uses the minimum value of cluster similarity over all clusters [34]. In this paper, we proposed a k-fold cross-validation procedure followed by the root mean squared error (RMSE) or the mean absolute percentage error (MAPE) to evaluate the prediction strength of the clustering algorithm. RMSE and MAPE are the most commonly used error measurements in statistics. In prediction tasks, RMSE indicates the absolute fit of the model to the data, i.e., it is used to compare how close the observed data points are to the predicted values of the model. MAPE is the average magnitude of the difference between predicted and actual values in percentages, without considering their direction, that is, since absolute percentage errors are used, the positive and negative errors are not canceling each other. In clustering validation, these two metrics can be used to measure the average distance between the data points and their cluster centers [35-37]. The smaller the RMSE/MAPE, the better the prediction results.

At each iteration of the k-fold CV procedure, one fold is used as the test set and the remaining folds as the training set. The training set is presented to a clustering method, giving a partition as a result (training partition). Then, new data points are assigned to the clusters in the training

partition based on the minimum distance from all the cluster centers. The CV method allows calculating the quality measure expressing the difference between the probability of occurrence of the outcomes (i.e., labels) in the training data and in the test data assigned to the same cluster. Once the clusters are formed using the training part of the data, the probability of occurrence of the labels in the training set, and in the testing set in each cluster will be assessed and analyzed. This is similar to estimating the probability that an outcome will occur, given that a sample belongs to a certain cluster, mathematically written as P(outcome|cluster). For instance, in the chronic disease dataset, one can estimate a probability of the risk of having hypertension in each of the generated clusters. Below, we describe the k-fold cross-validation procedure used to calculate a quality measure for a clustering model.

Let:

$L = \{ \lambda_i : i= 1,..., q \}$: the set of all labels in a multi-label dataset

$q = |L|$: the number of labels in the multi-label dataset.

k: the number of folds in the cross-validation procedure,

C: the number of clusters generated by the clustering algorithm.

Because we calculate label probabilities separately for each cluster i in each of the cross-validation folds j we denote these probabilities without using the number of the cluster nor the number of the fold in order not to clutter the equations:

y_m, m = 1,…, q: the probability that a sample from the training dataset assigned to cluster i has the m^{th} label

$\hat{y}_m$, m = 1,…, q: the probability that a sample from the testing dataset assigned to cluster i has the m^{th} label

1. Shuffle the original dataset randomly
2. Split the original dataset into k parts (folds) # k=10, for 10-fold cross-validation.
3. For each fold j=1,…,k.
    a) Take fold j as the test dataset (each fold, in turn, is used as the test dataset).
    b) Take the remaining folds together as the training dataset.
    c) apply dimensionality reduction (if needed)
    d) apply normalization to dataset (if needed)
    e) Generate clusters on the training dataset.
    f) Assign data points from the test dataset (selected in step ‘a’) into the corresponding clusters obtained in step ‘e.’

g) For each cluster i = 1, …, C found in step 'e':
    a. Compute the probabilities y_m, m = 1,…, q of the occurrence of the labels in cluster i based on the samples in the training dataset.
    b. Compute the probabilities $\hat{y}_m$, m = 1,…, q of the occurrence of the labels in cluster i using the assignment of the points from the test dataset to the clusters, which was obtained in step 'f.'
    c. Compute the root mean squared error ($RMSE_{ij}$) between the probabilities calculated in steps 'a.' and 'b.'. Note down the scores/errors as a quality measure for cluster i obtained in fold j.

4. When the loop in step 3 finishes (and so every fold served as the test set), take the average over the k folds of the recorded scores for each cluster and/or overall the clusters (equation (3)).

In the context of this study, RMSE and MAPE are proposed to measure the prediction strength of clustering techniques. RMSE represents the standard deviation of the difference between the probabilities of occurrence of the labels of the training data and the probabilities of occurrence of the labels of the test data in clusters. Intuitively, the RMSE in this study can be understood as the Euclidean distance between the vector of the observed probability scores of labels in the training data and the estimated probability scores of the labels in the test data for a given cluster, averaged by the total number of labels in the data (equation 1). Similarly, MAPE measures the size of the error between the probability scores of the training set and the probability scores of the test set in percentage terms (equation 2). RMSE and MAPE are evaluation methods that can be used together to diagnosis the variation in the errors of a clustering algorithm. For cluster i and cross-validation fold j these two measures are calculated as follows:

$$RMSE_{ij} = \sqrt{\frac{\sum_{i=1}^{q}(\hat{y}_i - y_i)^2}{q}} \quad (1)$$

$$MAPE_{ij} = \left(\frac{1}{q}\sum\frac{|y_i - \hat{y}_i|}{|y_i|}\right) * 100 \quad (2)$$

The resulting score obtained through RMSE with k-fold cross-validation across all clusters based on the probability score information from multiple labels, named CVIM in short, can be used as a cluster validity index (i.e., stability index). The better the values of the cluster validity index, the more stable the outputs of the clustering algorithm. High cluster stability is achieved when memberships of the clusters are not affected by small changes in the data set. The RMSE

of the clustering algorithm obtained using the k-fold cross-validation can be computed as shown in equation (3): let $RMSE_{ij}$ be the RMSE for the i^{th} cluster obtained in the j^{th} fold (equation 1). The average RMSE for the i^{th} clusters obtained in k fold with C clusters in each fold, denoted by $ARMSE_i$, can be computed as:

$$ARMSE_1 = (RMSE_{11} + RMSE_{12} + RMSE_{13} + \ldots + RMSE_{1k})/k$$
$$ARMSE_2 = (RMSE_{21} + RMSE_{22} + RMSE_{23} + \ldots + RMSE_{2k})/k$$
$$\vdots$$
$$ARMSE_C = (RMSE_{C1} + RMSE_{C2} + RMSE_{C3} + \ldots + RMSE_{Ck})/k$$

$$\text{Overall ARMSE} = (ARMSE_1 + ARMSE_{2+\ldots+} ARMSE_C)/C \quad (3)$$

$$\text{Cluster Validity Index(CVIM)} = \frac{1}{C}\sum_{i=1}^{C} ARMSE_i \quad (4)$$

Finally, the RMSE based cluster validity index across all clusters is found using equation (4). MAPE is also calculated in a similar fashion as the RMSE. The architecture of the proposed method for calculating RMSE and MAPE for each cluster in ten folds of cross-validation is presented in Figure 1 for an algorithm generating C = 3 clusters. In the final stage, the average RMSE/MAPE of 10 similar clusters is taken from each fold of cross-validation.

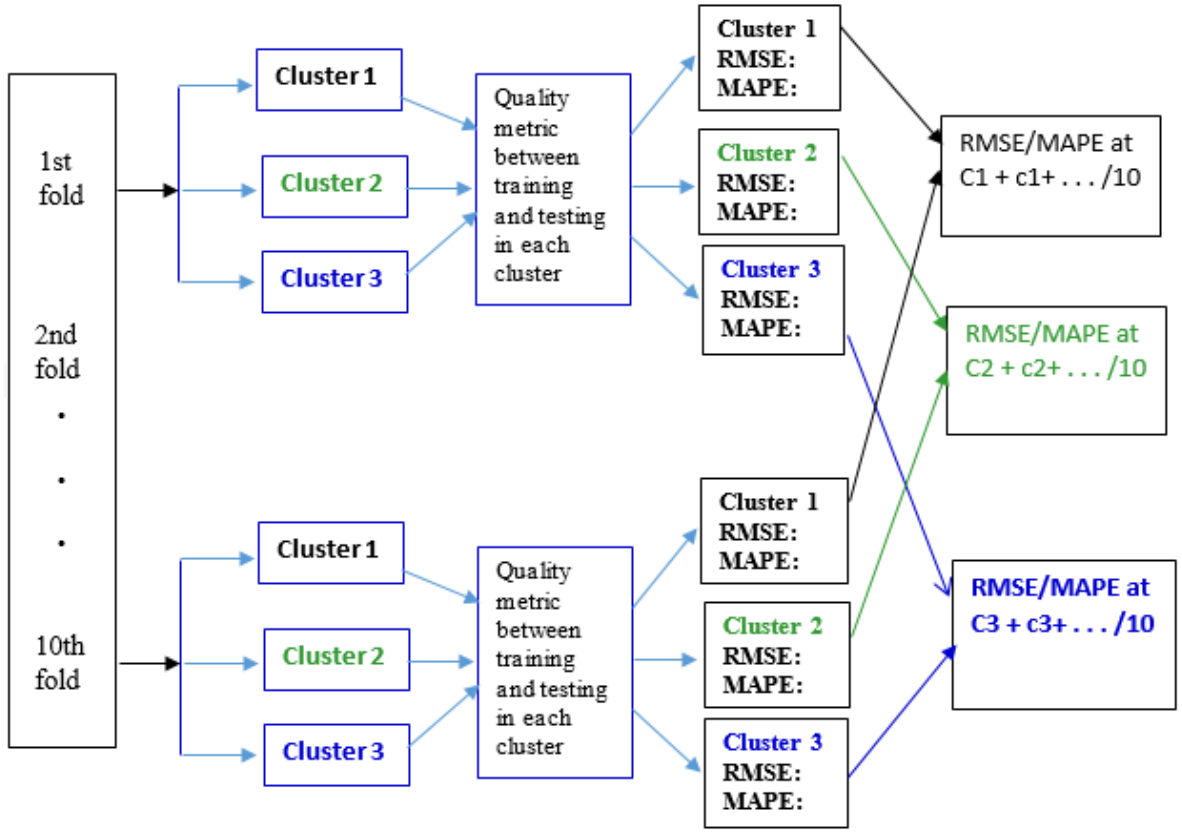


Figure 1. The architecture of the proposed method to evaluate a clustering model through 10 fold cross-validation with three clusters at each fold.

3. Experiments

In this paper, three public multi-label datasets were used to test the proposed method: the chronic diseases dataset [38], emotions [39], and Yeast [40] datasets. The chronic diseases dataset contains a collection of physical examination records for 110,300 patients with 62 features and 3 class labels. All the input features were used for forming clusters. The class labels (non-clustering variables), which include hypertension, diabetes, and fatty liver, were not used for defining clusters but only for cluster validation. Each record in the data may be associated with more than one of the class labels. As a result, the probability of occurrence of hypertension, diabetes, or fatty liver in patients of the test data can be estimated in the corresponding clusters. The chronic disease dataset is available online at http://pinfish.cs.usm.edu/dnn/. The Yeast dataset is formed by micro-array expression data and phylogenetic profiles with 2,417 genes. The dataset consists of 103 features with 14 labels, and each gene is associated with a set of functional labels. The emotions dataset contains examples

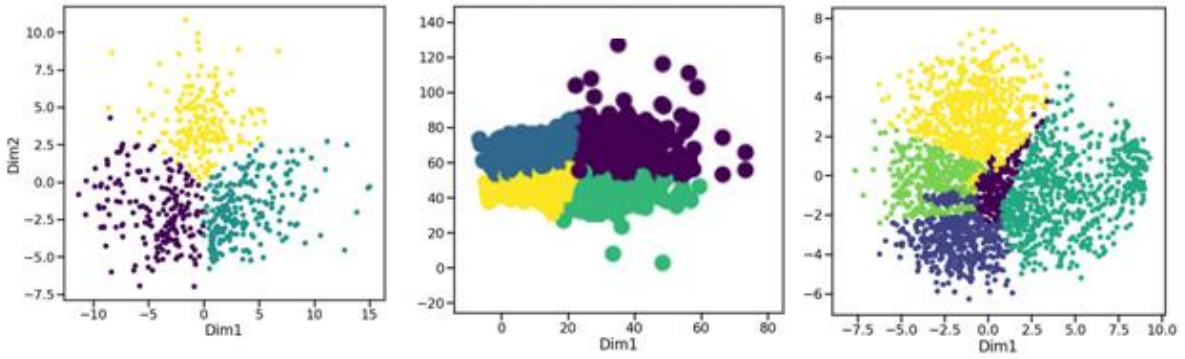
of songs according to people's emotions. The emotions and Yeast datasets were taken from the Mulan Library at http://mulan.sourceforge.net/datasets-mlc.html.

Multi-label datasets, and current data, in general, tend to be more complex than conventional data and need dimensionality reduction. All three multi-label datasets used in this experiment have a large number of features and labels/outcomes. Taking this problem into account, we applied the dimensionality reduction process to convert the dataset into two-dimensional space. The purpose of reducing data into lower-dimensional representation is to visualize and interpret the samples so that such visualization can be used to obtain insights from the data, e.g., to detect clusters and identify outliers. Moreover, a clustering process requires data reduction to obtain an efficient processing time while clustering and avoid the curse of dimensionality. For example, the k-means clustering algorithm often doesn't work well for high dimensional data [41]. There are different techniques proposed in the literature for high dimensional features in clustering [42,43]. In this study, principal component analysis (PCA) [44], one of the most commonly used technique, was applied as a data dimensionality reduction to convert each dataset into a two-dimensional representation. Emotions and Yeast datasets have large variations within the range of feature values, which can affect the quality of computed clusters. Therefore, after PCA, we applied the normalization technique [45] for Emotions and Yeast datasets to ensure that good quality clusters are generated. Then, k-means clustering [46] was applied to the reduced dataset. All the experiments have been implemented using Python programming language.

4. Results and Discussions

With the help of the Calinski-Harabasz index, three clusters for emotions dataset, four clusters for chronic disease dataset, and five clusters for yeast dataset were identified using the k-means clustering algorithm. A two-dimensional (2D) representation of clustering results for each dataset is shown in Figure 2. Colors of the points represent cluster memberships of the samples. For each dataset, the probabilities of the occurrence of each target variable in each cluster have been calculated both in the training and testing part of the data during the cross-validation procedure. We first evaluated the quality of the clusters using the existing internal validity criteria. Silhouette analysis is one of the most popular and effective internal measures which allows evaluating the appropriateness of the assignment of a data object to a cluster by measuring both intra-cluster cohesion and inter-cluster separation. Clusters within the range of 51 to 70% and 71 to 100% respectively indicate that a reasonable and a strong intra-cluster cohesion and inter-cluster separation are found [47]. The silhouette score can take values in the

interval [-1, 1]. Negative silhouette values represent wrong data placements, while positive silhouette values better data assignments. Therefore, we want the scores to be as big as possible and close to 1 to have good clusters. In our experiments, the silhouette score has shown good results. The silhouette score for clusters found on emotion, chronic disease, and Yeast datasets were 0.76, 0.82, and 0.69, respectively, indicating that the obtained clusterings were good ones.



(a) **(b)** **(c)**

Figure 2. 2D visualization of clustering results on Emotions (a), chronic disease (b), and Yeast (c) datasets. Min-Max normalization method has been applied to Emotions and Yeast datasets to eliminate the large variations within the range of features.

As the main objective of this study is to evaluate the prediction performance of the clustering algorithm through a 10-fold cross-validation procedure, the result of prediction performance in terms of RMSE and MAPE are presented for each cluster and across all clusters (i.e., the CVIM value), as shown in Table 1. The results represent the strength of the clustering algorithm to predict cluster labels for the test data. The obtained RMSE and MAPE scores of the clustering results in each cluster of each dataset represent the prediction errors.

Table 1. Performance of a clustering algorithm in each cluster and across the clusters (CVIM)

Dataset	**Metrics**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**CVIM***
Emotions	RMSE	0.021	0.019	0.017	-	-	0.019
	MAPE	7.88%	18.27 %	8.99 %	-	-	11.71%
Chronic	RMSE	0.0361	0.0543	0.0228	0.0282	-	0.0354
	MAPE	5.62%	5.92%	7.91%	12.29%	-	7.94%
Yeast	RMSE	0.071	0.061	0.066	0.086	0.076	0.072
	MAPE	7.49%	9.36 %	11.59 %	17.34 %	15.34 %	12.22%

CVIM*: RMSE/MAPE based cluster validity index across all the clusters in each of the three datasets.

Figures 3 and 4 show the RMSE and MAPE of the k-means clustering algorithm applied to each dataset, respectively. The smallest RMSE (i.e the better) is found in the Emotions dataset in each cluster, while the highest RMSE was found in the Yeast dataset. This also holds true

for the total RMSE across all the clusters (i.e., the CVIM score) on each dataset. Generally, an RMSE close to zero is indicative of the high similarity between the training and testing probabilities. Similarly, low MAPE values indicate good predictions of the occurrence of labels in each cluster across all datasets. The smaller the MAPE, the better the forecast, and more specifically, Lewis's [48] interpretation of MAPE is that a value of less than 10% indicates highly accurate forecast,11 to 20% is a good forecast, 21 to 50% is a reasonable forecast, and 51% or more is an inaccurate forecast. Accordingly, a highly accurate forecast is found in the chronic disease dataset. The results on emotion and yeast datasets show a good prediction.

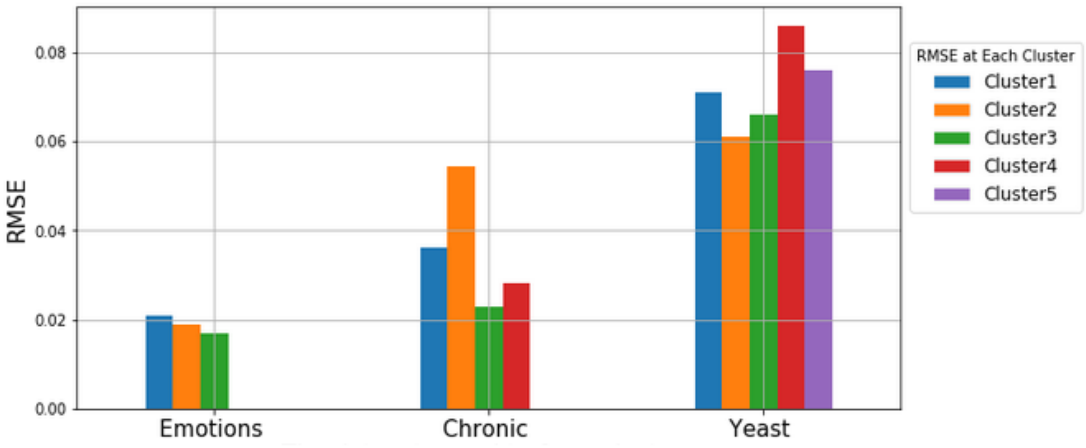


Fig.3. RMSE of the clustering algorithm on each cluster in each dataset

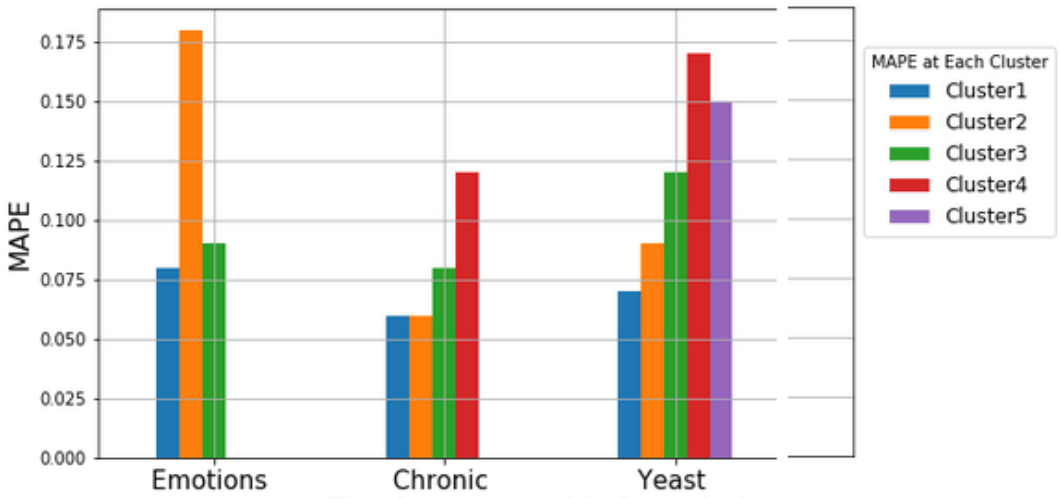


Fig 4. MAPE of the clustering algorithm on each cluster in each dataset

5. Conclusions

Evaluating the quality of clustering algorithms is an important and challenging part of the clustering task. In this study, the k-fold cross-validation procedure was adapted to the task of evaluating the quality of the clustering algorithms that is, measuring the ability of these algorithms to predict cluster membership for new data. A new clustering validity index was proposed to measure the effectiveness of the clustering algorithm through the use of root mean squared error (RMSE) and mean absolute percentage error (MAPE) values. The index was developed using the probability information obtained from several labels of multi-label data. This measure is useful for evaluating clusterings, which can be used for estimating the probability of the occurrence of the labels. For example, patients can be grouped into several clusters, and the occurrence of diseases can be studied separately in each group. The results presented in the paper show that the proposed method works well for evaluating the quality of clusters obtained using the k-means algorithm. Combining the proposed method with other, for example, density-based or hierarchical, clustering algorithms require solving additional problems such as finding an effective way of assigning new data points to previously discovered clusters. Therefore, combining the proposed method with such clustering algorithms was left as further work.

Compliance with Ethical Standards

Conflict of interest

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding

No funding was received for this study.

Acknowledgements

The authors would like to thank the reviewers of this paper for their supportive comments.

References

1. Wilks DS. Cluster Analysis. International Geophysics. 2011.

2. Xu R, WunschII D. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks. 2005;16:645–78.

3. Miljkovic D. Brief review of self-organizing maps. 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings. 2017.

4. Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35:1798–828.

5. Silverman BW. Density estimation: For statistics and data analysis. Density Estimation: For Statistics and Data Analysis. 2018.

6. Dokmanic I, Parhizkar R, Ranieri J, Vetterli M. Euclidean Distance Matrices: Essential theory, algorithms, and applications. IEEE Signal Processing Magazine. 2015;

7. Cordeiro De Amorim R, Mirkin B. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. Pattern Recognition. 2012;

8. Sahu L, Mohan BR. An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop. 9th International Conference on Industrial and Information Systems, ICIIS 2014. 2015.

9. Chakraborty S, Das S. K−Means clustering with a new divergence-based distance metric: Convergence and performance analysis. Pattern Recognition Letters. 2017;

10. Estivill-Castro V. Why so many clustering algorithms. ACM SIGKDD Explorations Newsletter. 2002;

11. Kleinberg J. An impossibility theorem for clustering. Advances in Neural Information Processing Systems. 2003.

12. Tanaka EA, Nozawa SR, Macedo AA, Baranauskas JA. A multi-label approach using binary relevance and decision trees applied to functional genomics. Journal of Biomedical Informatics. 2015;

13. Tarekegn A, Ricceri F, Costa G, Ferracin E, Giacobini M. Predictive Modeling for Frailty Conditions in Elderly People: Machine Learning Approaches. JMIR medical informatics. 2020;8:e16678. http://www.ncbi.nlm.nih.gov/pubmed/32442149

14. Tarekegn A, Ricceri F, Costa G, Ferracin E, Giacobini M. Detection of Frailty Using Genetic Programming. 23rd European Conference on Genetic Programming (EuroGP). Seville, Spain: Springer, Cham; 2020. p. 228–43. http://link.springer.com/10.1007/978-3-030-44094-7_15

15. Tan P-N, Steinbach M, Kumar V. Chap 8 : Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining. 2005;

16. Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association. 1971;

17. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research. 2010;

18. Rendón E, Abundez I, Arizmendi A, Quiroz EM. Internal versus External cluster validation indexes. International Journal of Computers and Communications. 2011;

19. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster

analysis. Journal of Computational and Applied Mathematics. 1987;20:53–65.

20. Caliñski T, Harabasz J. A Dendrite Method Foe Cluster Analysis. Communications in Statistics. 1974;

21. Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979;

22. Moulavi D, Jaskowiak PA, Campello RJGB, Zimek A, Sander J. Density-Based Clustering Validation. Proceedings of the 2014 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2014. p. 839–47.

23. Rakhlin A, Caponnetto A. Stability of K-means clustering. Advances in Neural Information Processing Systems. 2007.

24. Wang J. Consistent selection of the number of clusters via crossvalidation. Biometrika. 2010;

25. Tibshirani R, Walther G. Cluster validation by prediction strength. Journal of Computational and Graphical Statistics. 2005;

26. Ben-David S, Von Luxburg U. Relating clustering stability to properties of cluster boundaries. 21st Annual Conference on Learning Theory, COLT 2008. 2008.

27. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985;

28. Jain AK, Dubes RC. Clustering Methods and Algorithms. Algorithms for Clustering Data. 1988.

29. Saptarshi Chakraborty, Debolina Paul, Swagatam Das, Jason Xu: Entropy Weighted Power k-Means Clustering. AISTATS 2020: 691-701

30. Witten DM, Tibshirani R. A framework for feature selection in clustering. Journal of the American Statistical Association. 2010;

31. Olukanmi P, Nelwamondo F, Marwala T. Rethinking k-means clustering in the age of massive datasets: a constant-time approach. Neural Computing and Applications. 2019;

32. Zhang T, Ramakrishnan R, Livny M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. SIGMOD Record (ACM Special Interest Group on Management of Data). 1996;

33. Chakraborty S, Paul D, Das S. Hierarchical clustering with optimal transport. Statistics & Probability Letters. 2020;163:108781. https://linkinghub.elsevier.com/retrieve/pii/S0167715220300845

34. Hennig C, Meila M, Murtagh F, Rocci R. Handbook of cluster analysis. Handbook of Cluster Analysis. 2015.

35. Goran Petrović ŽĆ. Comparison of Clustering Methods for Failure Data Analysis: A Real Life Application. Proceedings of the XV International Scientific Conference on Industrial Systems (IS'11). 2011. p. 297–300.

36. Hassani M, Seidl T. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. Vietnam Journal of Computer Science. 2017;

37. Sidhu RS, Khullar S, Sandhu PS, Bedi RPS, Kaur K. A subtractive clustering based

approach for early prediction of fault proneness in software modules. World Academy of Science, Engineering and Technology. 2010;

38. Zhang X, Zhao H, Zhang S, Li R. A novel deep neural network model for multi-label chronic disease prediction. Frontiers in Genetics. 2019;

39. Trohidis K, Tsoumakas G, Kalliris G, Vlahavas I. Multi-label classification of music into emotions. ISMIR 2008 - 9th International Conference on Music Information Retrieval. 2008.

40. Elisseeff A, Weston J. A kernel method for multi-labelled classification. Advances in Neural Information Processing Systems 14. The MIT Press; 2002.

41. Napoleon D, Pavalakodi S. A New Method for Dimensionality Reduction Using KMeans Clustering Algorithm for High Dimensional Data Set. International Journal of Computer Applications. 2011;13:41–6.

42. Li W, Cerise JE, Yang Y, Han H. Application of t-SNE to human genetic data. Journal of Bioinformatics and Computational Biology. 2017;15:1750017.

43. Jin J, Wang W. Influential features PCA for high dimensional clustering. Annals of Statistics. 2016;

44. Syms C. Principal components analysis. Encyclopedia of Ecology. 2018.

45. Do JH, Choi DK. Normalization of microarray data: Single-labeled and dual-labeled arrays. Molecules and Cells. 2006.

46. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010;31:651–66.

47. Lv Y, Ma T, Tang M, Cao J, Tian Y, Al-Dhelaan A, et al. An efficient and scalable density-based clustering algorithm for datasets with complex structures. Neurocomputing. 2016;

48. Lewis. Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting. Butterworth Scientific. 1982;