**[[1]](#footnote-1) Transformed Data Obtained by Ensemble Clustering-Classification and Reduction**

Loai Abdallah and Malik Yousef

**Abstract**

The performance of many supervised or unsupervised machine learning algorithms depends very much on distance metrics to determine similarity between data points. A suitable distance metric could improve the classification performance, and clustering process significantly.

Distance metrics over a given range of data should reflect the actual similarity between objects. One of the obvious weaknesses of the Euclidean distance is dealing with data that is represented by a large number of attributes, where the Euclidean distance does not capture the actual relationship between those points. However, objects belonging to the same cluster usually share some common traits even though their Euclidean distance might be relatively large.

In this study, we propose a new classification method named *GrbClassifierEC* that replaces the given data space with categorical space based on ensemble clustering (EC). The similarity between objects is defined as the number of times that the objects belong to the cluster. The EC space is defined by tracking the membership of the points over multiple runs of clustering algorithms. Different points that were included in the same clusters will be represented as a single point. Our algorithm classifies all these points as a single class. In order to evaluate our suggested method, we compare its results to the *k* nearest neighbors, Decision tree and Random forest classification algorithms on several benchmark datasets. The results confirm that the suggested new algorithm *GrbClassifierEC* outperforms the other algorithms.

*Keywords*—Decision trees, Ensemble clustering, Classification.

# **INTRODUCTION**

T

his research presents a new classification model which classifies objects after running a mapping procedure that replaces a given data space with categorical space based on ensemble clustering (EC).

The main assumption in this research is that points belonging to the same cluster are more similar to other points from other clusters even though their Euclidean distance is closer. This is because the clustering algorithms take into account both the geometric space as well as other statistical parameters.

In this research we propose a procedure that transforms the original data space to another categorical feature space based on clustering algorithms. We call the new space EC space.

In general, the EC algorithm runs multiple clustering algorithms several times with different parameter values. Each data point is represented by the labels of the clusters it belongs to in each iteration yielding a categorical space. As a result, two different points may be represented identically if they were in the same clusters in each iteration. All the points that fall in the same cluster in the different clustering runs define an identical group and will be presented by a representor. Our algorithm classifies only the representors, and all the group members will have the same class label.

In our experiments we use the *k-means* clustering algorithm with different *k* values. We can see that not only the number of the data points (size) decreased, but also the number of features. This reduction is different than traditional feature reduction that eliminates some of the unneeded features. In the proposed new method we represent the data differently by clustering results.

Combination clustering is a more challenging task than the combination of supervised classifications. Topchy et al [1] and Strehl et al [2] addressed this issue by formulating consensus functions that avoid an explicit solution to the correspondence problem. Recent studies have demonstrated that consensus clustering can be found using graph-based, statistical or information-theoretic methods without explicitly solving the label correspondence problem as mentioned in [3]. Other empirical consensus functions were also considered in [4][5][6].

A clustering-based learning method was proposed in[7]. In this study, several clustering algorithms are run to generate several (unsupervised) models. The learner then utilizes the labeled data to guess labels for entire clusters (assuming that all points in the same cluster have the same label). In this way, the algorithm forms a number of hypotheses. The one that minimizes the PAC-Bayesian bound is chosen and used as the classifier. The authors assume that at least one of the clustering runs will produce a good classifier and that their algorithm will find it.

Ensemble clustering algorithms were applied also for semi-supervised classification[8][9] are based on the hypothesis is more accurately for noisy data to reflect the actual similarity between different objects. They propose a Co-association Matrix (CM) based on the outputs of different clustering algorithms and use this as a similarity matrix in the regularization framework.

Berikon et al[10] use the same idea in the semi-supervised regression method. They combine graph Laplacian regularization and cluster ensemble methodologies. To accelerate the calculation, they apply the low-rank decomposition of the CM.

Our method is different. We only assume that the groups, which were built by the identical points in the categorical space, are quite pure. Moreover, we do not integrate the clustering matrix with any classification algorithms; instead we classify the objects based on the groups’ classified members.

Abdallah et al [11][12] developed a distance function based on ensemble clustering and use it within the framework of the *k-nearest* neighbor classifier and then improve selecting sampling for unsupervised data to be labeled by an expert. Additionally Abddallah and Yousef [13] integrated EC within Decision Trees, K Nearest Neighbors, and the Random Forest classifiers. The results obtained by applying EC on 10 datasets confirmed the hypothesis that embedding the EC space would improve the performance and reduce the feature space dramatically.

A recent study by Yousef et al [14] used EC classification comparing it to two-class SVM and one-class classifiers applied on sequence plant microRNA data. The results show that K-Nearest Neighbors-EC (KNN-ECC) outperforms all other methods. The results emphasize that the EC procedure contributes to building a stronger model for classification.

Several experiments were conducted in order to evaluate the performance of the suggested method. We tested it over 10 datasets and compare its results to the *k nearest* neighbors, decision trees and random forest classification algorithms. The results show that the new algorithm using the ensemble clustering was superior and outperforms the other baseline algorithms on most of the datasets.

# **Ensemble Clustering Technique**

This section describes the ensemble clustering technique that we use in this research. The basic algorithm assumes that points belonging to the same cluster are more similar than points fall in different clusters. In real-world data, this assumption may not hold in each data. The following example illustrates this situation, in this example the data includes 2 classes (circles and diamonds). Suppose that we cluster this data into two clusters, we will get that the left cluster include two types of classes and the right one is pure (i.e., all the points from the same class).

To this end, we decided to run the clustering algorithm several times. Points belonging the same cluster in the multiple runs will define a and classified to same class.

## **The Ensemble Clustering Categorical Space**

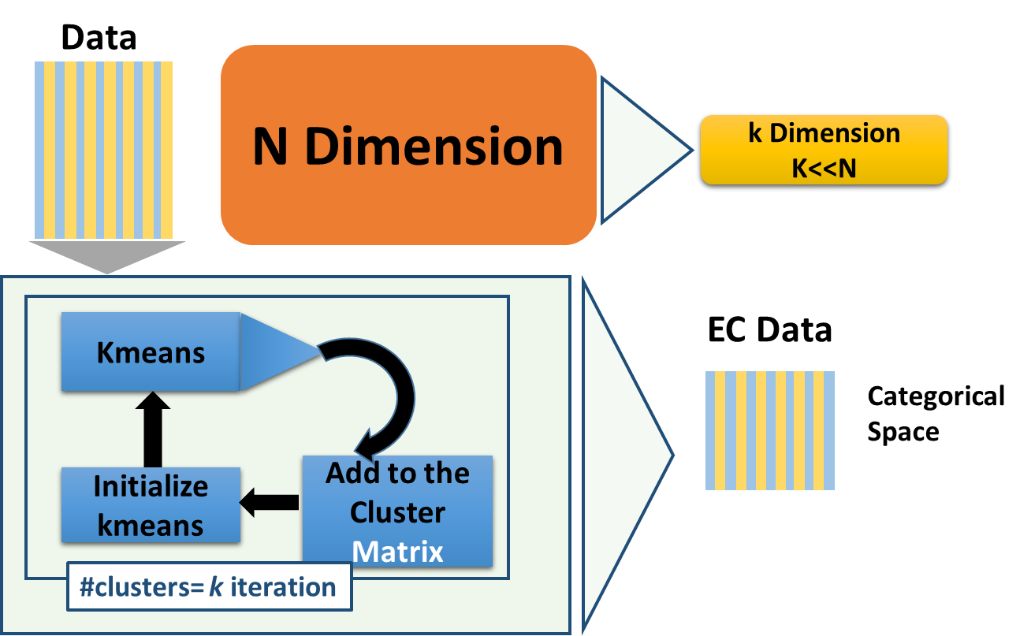
Here we describe how we transform the original data into the EC categorical space using the clustering method *k-*means. Let, be a set of labeled observations that used as a training data, A is a set of unlabeled data. Firstly, the algorithm will construct , where is a dataset of combining and (i.e., ), then the algorithm runs the k-means clustering algorithm several times with different values of (we refer it to *nmc* = number of clusters) and constructing the clustering matrix . is a matrix where the row consists of the clustering results of the object in . See Table 1 for an example.

At the end, each is transformed into new sample with categorical values. The dimension of the xi\* is *k.* We should emphasize that one need to consider categorical distance when applying similarity between two samples in the new categorical space. If in a specific run of k-means two samples or more have the same value then it means that they were put in the same cluster, otherwise they were put in different clusters. See Table 1 for example of 20 samples with *k*=11. We record the results from *k*=2 as with *k*=1 all the samples will be placed in one cluster.

Table 1: EC space for 20 samples and number of cluster (nmc) of 11. first column is the sample name, second column is the results of assigning k-means of each sample into two clusters (c0 and c1), the third column is the results of assigning k-means for each sample into 3 clusters and so on.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample/k** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** |
| **sample 1** | c0 | c2 | c3 | c2 | c2 | c4 | c5 | c4 | c4 | c5 |
| **sample 2** | c0 | c0 | c3 | c3 | c2 | c4 | c4 | c4 | c4 | c2 |
| **sample 3** | c0 | c2 | c2 | c4 | c5 | c5 | c6 | c6 | c6 | c6 |
| **sample 4** | c1 | c0 | c0 | c3 | c3 | c2 | c2 | c3 | c3 | c3 |
| **sample 5** | c0 | c0 | c3 | c3 | c2 | c2 | c4 | c2 | c2 | c2 |
| **sample 6** | c0 | c2 | c3 | c2 | c4 | c4 | c5 | c4 | c4 | c5 |
| **sample 7** | c0 | c2 | c3 | c2 | c4 | c4 | c5 | c5 | c5 | c4 |
| **sample 8** | c0 | c2 | c2 | c4 | c4 | c5 | c6 | c6 | c6 | c6 |
| **sample 9** | c1 | c0 | c0 | c3 | c3 | c2 | c2 | c3 | c3 | c3 |
| **sample 10** | c0 | c2 | c3 | c2 | c4 | c4 | c5 | c5 | c4 | c5 |
| **sample 11** | c0 | c2 | c2 | c2 | c4 | c5 | c6 | c5 | c5 | c4 |
| **sample 12** | c0 | c2 | c2 | c2 | c4 | c5 | c6 | c5 | c5 | c4 |
| **sample 13** | c0 | c2 | c2 | c2 | c4 | c5 | c6 | c5 | c5 | c4 |
| **sample 14** | c0 | c2 | c3 | c2 | c2 | c4 | c5 | c4 | c4 | c5 |
| **sample 15** | c0 | c2 | c2 | c2 | c4 | c5 | c6 | c5 | c5 | c4 |
| **sample 16** | c0 | c2 | c3 | c2 | c4 | c4 | c5 | c5 | c4 | c5 |
| **sample 17** | c0 | c2 | c3 | c2 | c4 | c5 | c5 | c5 | c5 | c4 |
| **sample 18** | c0 | c2 | c3 | c2 | c2 | c4 | c5 | c4 | c4 | c5 |
| **sample 19** | c0 | c0 | c3 | c3 | c2 | c2 | c4 | c2 | c2 | c2 |
| **sample 20** | c0 | c2 | c2 | c2 | c4 | c5 | c6 | c5 | c5 | c4 |

Figure 1: The workflow for creating the EC categorical space based on k-means clustering algorithm. The original data is the input to the workflow. The outcome is a new data named EC data in a categorical space with dimension k. the sign << indicates that k is dramatically smaller than the original data dimension N.



***EC Transformation***

**Input:**

E(*l*,N) : *x1 , x2,…,xl* Data consists of *l* samples in N dimension(features)

*k*: number of clusters

create empty matrix *cMat* with *l* rows (number of samples) and *k* columns.

**Algorithm:**

For each *nmc in {1,2,3,…,k}* do*:*

cMat{:,*nmc*} = k-means(E, *nmc*); assign for each sample *xi* a cluster c0,c1,..,ck-1

(see Table 1 for an example of cMat)

Algorithm 1:EC transformation algorithm. The k-means used as the clustering algorithm.

## **Reduction of the EC sample**

The new categorical data that result of applying the EC transformation (*Algorithm 1*), consists of *l* samples with *k* categorical features. The feature space is reduced dramatically and now the new dimension *k* is much less that the original data dimension (k<<N in Figure 1). More interestingly, the new EC data could be reduced in term of sample dimension. Samples or points that share the same cluster all over the *k* iteration of *k*-means are consider to be one point. For example, considering Table 1, sample11, sample 12 and sample 20 have the same categorical values. The vector space that represent those 3 points is =(c0, c2, c2, c2, c4, c5, c6, c5, c5, c4), additionally the points (samples) sample 1 and sample 18 have the same values then can be represented and reduced to one point. Therefore, the new EC samples are redundant and can be presented based on the representors .

We have iterated all over the points in the EC data and keep the representor for each group.

Note that, the set contains labeled and unlabeled data, as a result the groups may contain labeled and unlabeled objects. Generally, there are four possible cases for the objects that were grouped together:

1. All the objects are classified as a same class: in this case the group also will be classified as the class of its objects.
2. All the objects are classified but their classes are different: Then the group will be classified as the majority class.
3. Some of the objects are classified and the rest are not: the same like (2).
4. All the objects are not labeled: in this case the group will be unclassified group.

To this end, we define a purity measurement for a group in order to evaluate the grouping process. The purity measurement based mainly of the probabilities of the labeled objects as follow:

where denotes group that was represented by vector in the matrix , denotes the number of the members , and denotes the probability of class in group . As can be seen, equal 1 when the group is pure and for the lowest purity, that will decrease as the number of the classes increases.

# **Ensemble Clustering Based Classifier**

In this section we describe our new classifier approach named GrbClassifierEC. The pseudo code of the algorithm is presented in *Algorithm 2*. The main idea of the classifier is to generate the EC unique samples from the generated EC samples, which actually is the representative set of EC samples. then we need to check for each represented EC sample the distribution of the labels in its original group.

***Grouping based classifier***

**Input:**

a matrix with the ensemble clustering results.

E(*l*,N) : *x1 , x2,…,xl* Data consists of *l* samples in N dimension(features)

*k*: number of clusters

create empty matrix *cMat* with *l* rows (number of samples) and *k* columns.

**Algorithm:**

1. Create the based on the EC results.
2. For each :
   1. Repeat until stopping criteria satisfies:
      1. Select labeled representor .
      2. Assign the label of to all the unlabeled members.
   2. Classify all the unlabeled members by the majority class that they have.
   3. Calculate the
   4. The accuracy for each unlabeled member will be as the group purity.
3. Return the labeled dataset.

d

Algorithm 2 : Grouping EC based Classifier named GrbClassifierEC is our new approach for classification based EC

# **Experiments on Numerical Datasets**

To evaluate the merit of the new classifier GrbClassifierEC we compare its results to the k-nearest neighbors, decision trees and random forest classification algorithms. We tested it over 10 datasets and we compare the performance for each algorithm. The results shown that the new algorithm using the ensemble clustering was superior and outperform the other baseline algorithms on most the datasets.

# **Datasets**

The data is consisting of microRNA precursor sequences where each sequence consist of 4 nucleotide letters {A,U,C,G,} and the length of each precursor sequence is about 70 nucleotides. The source of such data is miRbase[15]. We have used part of the data that was used in different studies[16,17] and also used in our previouse study [13].

One simple way of representing sequences that consist of 4 nucleotide letters is by employing k-mers frequency. The -mer counts in a given sequence were normalized by the length of the sequence.

Our features are including k-mer frequencies, other distance features that just recently was suggested by Yousef et al (2019) (still not published), secondary features suggested suggest by [18] and additionally many features describing pre-miRNAs have been proposed [19] are included in the features set. The number of features is 1038 features.

The main data consists of information from 15 clades(Table 2). The sequences of *Homo sapiens* were taken out of the data of its clade Hominidae. The datasets were passed a process of removing homology sequences (keeping just one representative). One can generate about 256 datasets by considering a pair of two clades including itself. We have considered randomly 10 datasets from those set of datasets listed in Table 3.

Table 2: The list of clades used in the study, the first column presents the name of the clade, the second column the number of pre-cursors available on miRBase, and the third column is a number of precursors after preprocessing the data.

|  |  |  |
| --- | --- | --- |
| Data set | Number of Precursors | Number of Unique Precursors |
| Hominidae | 3629 | 1326 |
| Brassicaceae | 726 | 535 |
| Hexapoda | 3119 | 2050 |
| Monocotyledons (Liliopsida) | 1598 | 1402 |
| Nematoda | 1789 | 1632 |
| Fabaceae | 1313 | 1011 |
| Pisces (Chondricthyes) | 1530 | 682 |
| Virus | 306 | 295 |
| Aves | 948 | 790 |
| Laurasiatheria | 1205 | 675 |
| Rodentia | 1778 | 993 |
| *Homo sapiens* | 1828 | 1223 |
| Cercopithecidae | 631 | 503 |
| Embryophyta | 287 | 278 |
| Malvaceae | 458 | 419 |
| Platyhelminthes | 424 | 381 |

Table 3: 10 datasets. The first column is the name of the first clade positive data, second column is the second clade negative data.

|  |  |
| --- | --- |
| Positive Data | Negative Data |
| Aves | Embryophyta |
| Cercopithecidae | Malvaceae |
| Embryophyta | Laurasiatheria |
| Fabaceae | Nematoda |
| Hexapoda | Aves |
| Laurasiatheria | brassicaceae |
| Malvaceae | Fabaceae |
| brassicaceae | Hexapoda |
| hominidae | Cercopithecidae |
| Monocotyledons | homoSapiens |

# **Reduction of the EC SAmple**

For each unique point we have measure its size, the size here is the number of times this unique point appears in the EC data. For example, see Table 3, we have 305 unique points with size 1, that’s means all those 305 points appear once in the data, while we see 68 unique points that each one appears twice in the data, then its size is 2. We have 22 points with size 3, that means each points of the 22 unique point appears 3 times in the data. We should indicate that the labels are not included in the EC data. That’s mean that group of points at the EC space can have different labels associated to the original points and still share the same group.

Table 3 demonstrate the output of the EC procedure with *k*=30 applied on the data Cercopithecidae vs Malvacea that contains 894 examples (points). Table 3 shows that the EC data has 449 unique points which is 50% reduction in the size of the original data (449/894=0.5).

Table 3: The data Cercopithecidae vs Malvacea with k=30. The total number of samples (points) is 894 which is the sum of column #Points. The size of the unique points is the sum of columns “Unique Points” which is 449.#Points is multiplication of Size and Unique Points. Ratio Unique Points is the #Unique Points/Total #Points while Ratio All is #Points/Total #Points.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Size | Unique Points | #Points | Ratio Unique Points | Ratio All |
| 1 | 305 | 305 | 67.929% | 34.116% |
| 2 | 68 | 136 | 30.290% | 15.213% |
| 3 | 22 | 66 | 14.699% | 7.383% |
| 4 | 18 | 72 | 16.036% | 8.054% |
| 5 | 11 | 55 | 12.249% | 6.152% |
| 6 | 5 | 30 | 6.682% | 3.356% |
| 7 | 5 | 35 | 7.795% | 3.915% |
| 10 | 4 | 40 | 8.909% | 4.474% |
| 13 | 3 | 39 | 8.686% | 4.362% |
| 8 | 3 | 24 | 5.345% | 2.685% |
| 9 | 2 | 18 | 4.009% | 2.013% |
| 29 | 1 | 29 | 6.459% | 3.244% |
| 14 | 1 | 14 | 3.118% | 1.566% |
| 31 | 1 | 31 | 6.904% | 3.468% |
| Total | **449** | **894** |  |  |

Figure 2 is presents the distribution of the group size for *k*=30 and *k*=50. It is clear that as the *k* is increasing the number of groups with size 1 is increasing. One expect to get number of groups of size of 1 to be as the number of the original number of sample as we increasing the value of k. In other words, each sample will be hosted in one cluster. This actually raise a scientific question, what is the optimal value of *k* that will yield in improving the performance of the classifier or more specifically capturing the nature of the data in terms of clusters.

Figure :Distributaion of the groups samples (points) size comparing nmc=30 and nmc=50.

## **Model Performance Evaluation**

We have tested a different number of EC clusters ranging from 10 to 100 iterated 10 times. For each level, we have run 100 iterations with equal sample size and calculate the mean of each performance measurements described below.

For each established model, we calculated a number of performance measures for the evaluation of the classifier such as sensitivity, specificity, and accuracy according to the following formulations (with TP: true positive, FP: false positive, TN: true negative, and FN referring to false negative classifications):

## **Results**

We have conducted a comparison study for the new classifier GrbClassifierEC with the other known classifiers such as k-nearest neighbors, decision trees and random forest classifiers. The results are presented in Table 5. The results are clearly showing that the performance of the suggested classifier GrbClassifierEC was superior.

Figure 3 shows the performance of different classifier over different level of training percentage of the data. The results of EC are referring to our GrbClassifierEC classifier. We see that the performance is not dramatically influences by the size of the training part for the other classifiers while it increasing dramatically for GrbClassifierEC classifier the 39% level. Moreover, it could reach a very high performance as the perchance of the training part is increasing which is actually a function of the value of k in the EC transformation.

In term of data reduction, Table 5 and Table 6 demonstrate that about 56% of the samples data are reduced in the EC space with *k* of 49 and 39% in the EC space with *k* of 30. Those results demonstrate the advantage of our approach in reducing the data size and could be a contribution to be used for big data.

Table 5 and Table 6 shows the comparison results of the EC classifier with other classifiers applied on the whole feature space (named Regular Classifiers) and the performance of Random forest applied on the EC categorical data(EC-RF).

Figure :The accuracy of the classifiers over different level of sample training size.

Table 5 present results with *k* of 49 while table 6 present results with *k* 3. Interestingly, EC Classifier outperforms all the other approaches while using just 56% in average of the data (see ratio column) while the regular classifiers using 80% of the data for training. EC classifier is outperforming the regular approaches by 9% for the DT,6% for the KNN, 8% for the random forest applied on the EC sample and by 3% for the regular random forest.

Table 5: GrbClassifierEC -EC classifier results with k of 49 compared to Random forest applied on the EC samples and results for regular classifiers applied on the original data. K is number of clusters

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data/Performance** | **Data Info** | | | **EC Classifier**  GrbClassifierEC | | | | **Acuuracy Diffrence** | | | | **EC-RF** | | | **Regular Classifiers** | | |
| **#Sample** | **#EC\_Samples** | **ratio** | **Sensitivity** | **Specifity** | **F-measure** | **Accuracy** | **EC Random Forest** | **Random Forest** | **DTT** | **KNN** | **Sensitivity** | **Specifity** | **Accuracy** | **AccDT** | **AccKNN** | **AccRF** |
| Aves vs Embryophyta | 1068 | 726 | 68% | 0.97 | 0.92 | 0.97 | 0.96 | 0.02 | 0.01 | 0.05 | 0.02 | 0.84 | 0.97 | 0.93 | 0.91 | 0.93 | 0.95 |
| Cercopithecidae vs Malvaceae | 894 | 593 | 66% | 0.98 | 0.97 | 0.98 | 0.98 | 0.08 | 0.05 | 0.10 | 0.07 | 0.84 | 0.94 | 0.90 | 0.88 | 0.91 | 0.93 |
| Embryophyta vs Laurasiatheria | 953 | 652 | 68% | 0.96 | 0.92 | 0.96 | 0.95 | 0.08 | 0.04 | 0.10 | 0.07 | 0.94 | 0.72 | 0.87 | 0.85 | 0.88 | 0.91 |
| Fabaceae vs Nematoda | 2642 | 1004 | 38% | 0.85 | 0.89 | 0.84 | 0.87 | 0.02 | -0.01 | 0.04 | 0.00 | 0.92 | 0.76 | 0.85 | 0.83 | 0.88 | 0.89 |
| Hexapoda vs Aves | 2840 | 2087 | 73% | 0.85 | 0.95 | 0.86 | 0.92 | 0.10 | 0.03 | 0.11 | 0.10 | 0.61 | 0.91 | 0.83 | 0.81 | 0.82 | 0.89 |
| Laurasiatheria vs Brassicaceae | 1209 | 570 | 47% | 0.93 | 0.93 | 0.94 | 0.93 | 0.05 | 0.01 | 0.05 | 0.02 | 0.86 | 0.90 | 0.88 | 0.89 | 0.91 | 0.92 |
| Malvaceae vs Fabaceae | 1401 | 749 | 53% | 0.69 | 0.87 | 0.68 | 0.82 | 0.16 | 0.05 | 0.15 | 0.12 | 0.84 | 0.22 | 0.67 | 0.67 | 0.70 | 0.77 |
| brassicaceae vs Hexapoda | 2584 | 870 | 34% | 0.84 | 0.96 | 0.84 | 0.93 | 0.02 | 0.00 | 0.03 | 0.01 | 0.97 | 0.74 | 0.92 | 0.90 | 0.93 | 0.94 |
| Hominidae vs Cercopithecidae | 1829 | 1059 | 58% | 0.72 | 0.91 | 0.73 | 0.86 | 0.15 | 0.09 | 0.20 | 0.14 | 0.25 | 0.87 | 0.70 | 0.66 | 0.71 | 0.76 |
| Monocotyledons vs HomoSapiens | 2625 | 1460 | 56% | 0.92 | 0.93 | 0.92 | 0.92 | 0.10 | 0.03 | 0.09 | 0.04 | 0.84 | 0.82 | 0.83 | 0.83 | 0.88 | 0.89 |
| **Average** |  |  | **56%** | **87%** | **92%** | **87%** | **91%** | **8%** | **3%** | **9%** | **6%** | **79%** | **78%** | **84%** | **82%** | **85%** | **89%** |

Table 6 demonstrate that one ca reduces more the size of the data to reach 39% ration with *k*=30 and still get a reasonable result. The EC classifier outperform DTT and EC-RF and KNN with 5%,3% and 1% respectively, while RF outperforms it with 2%. More interesting that ration of the reduction is an indication about the data redundant and the similarity of the original points of the data.

Table 6: GrbClassifierEC - EC classifier results with k of 30 compared to Random forest applied on the EC samples and results for regular classifiers applied on the original data. K is number of clusters. The section “Accuracy Difference” is EC Classifier-ACC of the other classifier. A value of positive means EC classifier is better than the corresponding classifier. EC-RF is random forest applied on the EC data, RF is random forest applied on the original data. DTT is decision tress while KNN is K- Nearest Neighbors applied on the original data.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data/Performance** | **Data Info** | | | **EC Classifier**  GrbClassifierEC | | | | **Acuuracy Diffrence** | | | |
| **#Sample** | **#EC\_Samples** | **ratio** | **Sensitivity** | **Specifity** | **F-measure** | **Accuracy** | **EC-RF** | **RF** | **DTT** | **KNN** |
| **Aves vs Embryophyta** | 1068 | 513 | 48% | 0.86 | 0.94 | 0.85 | 0.92 | -0.01 | -0.03 | 0.02 | -0.01 |
| **Cercopithecidae vs Malvaceae** | 894 | 449 | 50% | 0.94 | 0.92 | 0.94 | 0.94 | 0.04 | 0.01 | 0.06 | 0.03 |
| **Embryophyta vs Laurasiatheria** | 953 | 493 | 52% | 0.94 | 0.83 | 0.94 | 0.91 | 0.04 | 0.00 | 0.06 | 0.03 |
| **Fabaceae vs Nematoda** | 2642 | 536 | 20% | 0.78 | 0.88 | 0.79 | 0.84 | -0.01 | -0.05 | 0.01 | -0.04 |
| **Hexapoda vs Aves** | 2840 | 1647 | 58% | 0.76 | 0.92 | 0.78 | 0.88 | 0.05 | -0.01 | 0.07 | 0.06 |
| **Laurasiatheria vs Brassicaceae** | 1209 | 406 | 34% | 0.89 | 0.88 | 0.89 | 0.88 | 0.00 | -0.04 | 0.00 | -0.03 |
| **Malvaceae vs Fabaceae** | 1401 | 451 | 32% | 0.55 | 0.80 | 0.53 | 0.73 | 0.07 | -0.04 | 0.06 | 0.03 |
| **brassicaceae vs Hexapoda** | 2584 | 542 | 21% | 0.77 | 0.95 | 0.78 | 0.91 | -0.01 | -0.03 | 0.01 | -0.02 |
| **Hominidae vs Cercopithecidae** | 1829 | 786 | 43% | 0.61 | 0.87 | 0.63 | 0.80 | 0.10 | 0.04 | 0.14 | 0.09 |
| **Monocotyledons vs HomoSapiens** | 2625 | 855 | 33% | 0.86 | 0.87 | 0.86 | 0.87 | 0.04 | -0.03 | 0.03 | -0.01 |
| **Average** |  |  | **39%** | **80%** | **89%** | **80%** | **87%** | **3%** | **-2%** | **5%** | **1%** |

# **Conclusion**

In this work we have demonstrated the advantage of the EC approach in reducing the feature space and also in reducing the data size. Additionally, we have proposed a new classifier approach named GrbClassifierEC based on the EC data. Generally speaking, we shown that we are able to reduce the number of features dramatically to be 5% or 3% (50/1038 = 0.048, 30/1038=0.0.28) and reduce the size of the data to 56% and 39% and still get similar performance to regular classifier applied on the original data or even in some cases outperform them. However, those results are obtained in a pay off in computation time that the ES transformation algorithm requires.

The main assumption was, points within the same cluster share common traits more than points within different clusters. Thus, representing the objects based on the clustering space it may be better than the geometric space.

The approach suggested here is very useful for the field of big data that allowed to reduce the data to a representative data by considering its EC data. As a future work we will need to suggest and algorithm that would pick the optimal value of k that would yield in improving the performance under the constrains of reducing the size of the data dramatically.

Our algorithm, however, is general and can be integrated with many algorithms. In this research, we use only the k-means clustering algorithm with different k values. In the future work, there are several directions: (1) checking the effect of the clustering algorithm to build an ensemble clustering space. (2) how to detect poor clustering results based on the training data. (3) how to reduce the

the volume of the data by combine similar point based on the EC.

**Acknowledgment**

This research was supported by the Max Stern Yezreel Valley College for LA and by zefat ACADEMIC college for my.

**References**

1. Topchy a., Jain a. K, Punch W. Combining multiple weak clusterings. Third IEEE Int Conf Data Min. 2003;0–7.

2. Strehl A, Ghosh J. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. J Mach Learn Res. 2002;3:583–617.

3. Topchy A, Jain AK, Punch W. Clustering ensembles: Models of consensus and weak partitions. IEEE Trans Pattern Anal Mach Intell. 2005;27:1866–81.

4. Dudoit S, Fridlyand J. Bagging to improve the accuracy of a clustering procedure. Bioinformatics [Internet]. 2003;19:1090–9. Available from: http://dx.doi.org/10.1093/bioinformatics/btg038

5. Fern XZ, Brodley CE. Random projection for high dimensional data clustering: A cluster ensemble approach. Proc Twent Int Conf Mach Learn [Internet]. 2003;20:186–93. Available from: http://www.aaai.org/Papers/ICML/2003/ICML03-027.pdf

6. Fischer B, Buhmann JM. Bagging for path-based clustering. IEEE Trans Pattern Anal Mach Intell. 2003;25:1411–5.

7. Derbeko P, El-Yaniv R, Meir R. Explicit learning curves for transduction and application to clustering and compression algorithms. J Artif Intell Res. 2004;22:117–42.

8. Berikov V, Karaev N, Tewari A. Semi-supervised classification with cluster ensemble. Proc - 2017 Int Multi-Conference Eng Comput Inf Sci Sib 2017. 2017.

9. Yu GX, Feng L, Yao GJ, Wang J. Semi-supervised classification using multiple clusterings. Pattern Recognit Image Anal [Internet]. 2016;26:681–7. Available from: https://doi.org/10.1134/S1054661816040210

10. Berikov V, Litvinenko A. Semi-Supervised Regression using Cluster Ensemble and Low-Rank Co-Association Matrix Decomposition under Uncertainties. 2019 [cited 2019 Mar 4]; Available from: http://arxiv.org/abs/1901.03919

11. AbedAllah L, Shimshoni I. k Nearest Neighbor Using Ensemble Clustering. In: Cuzzocrea A, Dayal U, editors. Data Warehous Knowl Discov 14th Int Conf DaWaK 2012, Vienna, Austria, Sept 3-6, 2012 Proc [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 265–78. Available from: http://dx.doi.org/10.1007/978-3-642-32584-7\_22

12. AbdAllah L, Shimshoni I. An ensemble-clustering-based distance metric and its applications. Int J Bus Intell Data Min [Internet]. 2013;8:264–87. Available from: http://www.inderscienceonline.com/doi/abs/10.1504/IJBIDM.2013.059052

13. Abddallah L, Yousef M. Ensemble Clustering Based Dimensional Reduction. In: Elloumi M, Granitzer M, Hameurlain A, Seifert C, Stein B, Tjoa AM, et al., editors. Database Expert Syst Appl. Cham: Springer International Publishing; 2018. p. 115–25.

14. Yousef M, Khalifa W, AbedAllah L. Ensemble Clustering Classification compete SVM and One-Class classifiers applied on plant microRNAs Data. J Integr Bioinform. Germany; 2016;13:304.

15. Griffiths-Jones S. miRBase: microRNA sequences and annotation. Curr Protoc Bioinformatics. 2010;Chapter 12:Unit 12.9.1-10.

16. Yousef M, Nigatu D, Levy D, Allmer J, Henkel and W. Categorization of Species based on their MicroRNAs Employing Sequence Motifs, Infor-mation-Theoretic Sequence Feature Extraction, and k-mers. EURASIP J Adv Signal Process. 2017;

17. Yousef M, Khalifa W, Acar \.Ilhan Erkin, Allmer J. MicroRNA categorization using sequence motifs and k-mers. BMC Bioinformatics [Internet]. 2017;18:170. Available from: http://dx.doi.org/10.1186/s12859-017-1584-1

18. Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. Bioinformatics [Internet]. 2006;22:1325–34. Available from: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/11/1325

19. Sacar MD, Allmer J. Data mining for microrna gene prediction: On the impact of class imbalance and feature number for microrna gene prediction. 2013 8th Int Symp Heal Informatics Bioinforma. IEEE; 2013. p. 1–6.

**Loai A. Abdallah** received his B.Sc. in Mathematics and Management Information Systems from the University of Haifa, his M.Sc. and Ph.D. in Mathematics from the University of Haifa. Loai was a member of the Departments of Mathematics and Computer Science at the College of Sakhnin from October 2011. He joined the department of Community Information Systems at Zefat academic college from October 2011. Currently, Loai is a member in the department of Information Systems from October 2016 in the Max Stern Yezreel Valley College.

Dr. Abdallah is active in the industry. He is a co- founder and the Chief technology officer in iDRiSi Company.

**Malik Yousef** is a data scientist, with focus on bioinformatics with applications to various biomedical/biological problems.  He has published more than 55 peer‐reviewed articles in top journals and proceedings with over 2400 citations and an H-index of 18 and i10-index of 20 (based on Google scholar.

His international experience includes 3 years as a postdoc at The Wistar Institute, Cancer Center, USA [Prof Louise Showe Cancer Biology lab] and one year at the University of Pennsylvania [UPENN-Bioinformatics Center]. Currently he is Assistant Professor at the Zefat Academic College in Israel.

1. L. A. Loai is with the Department of Information Systems, The Max Stern Yezreel Valley Academic College, Israel. (corresponding author, phone: +972 (505) 714 178; e-mail: Loai1984@gmail.com).

   Y. Malik, the Department of Community Information Systems, Zefat Academic College, Zefat, 13206, Israel; e-mail: malik.yousef@gmail.com [↑](#footnote-ref-1)