Transformed Data Obtained by Ensemble Clustering-Classification and Reduction

Loai Abdallah and Malik Yousef

Abstract

The ~~P~~performance of many ~~machine learning algorithms~~ supervised or unsupervised machine learning algorithms ~~are~~ depends ~~critically~~ very much on distance metrics to determine similarity between data points. A suitable distance metric ~~might be the cause in~~could improvinge the ~~performance of~~ classification performance, and clustering process significantly.

Distance metrics over a given ~~space~~ range of data should reflect~~s~~ 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 replace~~d~~s the given data space ~~to~~ with categorical space based on ensemble clustering (EC). ~~and t~~The similarity between objects is defined as the number of times that the objects ~~were~~ belong~~ing~~ 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 ~~same~~single point. Our algorithm classifies all these points as a single~~same~~ class(**we mean we assign those points to be belongs to one class, mainly we have two-class data**). 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.

Comment [A1]: What is the meaning of k here?

Comment [A2]: It is part of the algorithm name

Formatted: Font: 12 pt, Italic, Complex Script Font: 12 pt, Italic

Formatted: Default Paragraph Font, Font: Italic, Complex Script Font: Italic

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

I. INTRODUCTION

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

The main assumption in this research is that points ~~that~~ 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 ~~consider not only~~take into account both the geometric space ~~but also~~as well as other statistical parameters.

In this research we propose a ~~transformation~~ 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 ~~will be~~is represented by the labels of the clusters it ~~was~~ belongs~~ing~~ 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.~~,~~ ~~a~~All the points that fall~~s~~ in the same cluster in the different

Comment [A3]: Do you mean represented?

Comment [A4]: yes

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

clustering runs ~~will~~ 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~~amount~~ of the data points (size) ~~was~~ decreased, but also the number of ~~the~~ features~~ also is decreased~~. This reduction is different than ~~the~~ traditional feature reduction~~,~~ that eliminates some of the unneeded features.~~,~~ ~~i~~In the proposed~~s~~ new method we represent the data ~~simply~~ differently by ~~the~~ 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 (~~under the assumption~~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 C~~c~~o-association M~~m~~atrix (CM) based on the outputs of different clustering algorithms ~~runs~~ and use this~~it~~ 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 ~~differing from all those works~~different. We only assume ~~only~~ 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,~~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 ~~they~~ 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 ~~hypotheses~~ hypothesis that embedding the EC space would improve the performance and reduce the feature space dramatically.

A recent study by Yousef et al [14] ~~has~~ 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~~n~~ that the new algorithm using the ensemble clustering was superior and outperforms the other baseline algorithms on most of the datasets.

Comment [A5]: What do you mean here?

Comment [A6]: One point that represent all the points belongs to the group. We mean like a person who represent a group of people.

Formatted: Font: Italic, Complex Script Font: Italic

Comment [A7]: Not clear to me whether this is a new sentence or not.

Comment [A8]: Do you mean boundary?

Comment [A9]: This sentence is unclear, please clarify.

Comment [A10]: Do you mean: Our method works differently?

Comment [A11]: We want to say that our approach is not similar to other approaches already published

Comment [A12]: Do you mean relatively accurate?

Comment [A13]: No- it is from purity

Formatted: Font: Italic, Complex Script Font: Italic, Check spelling and grammar

Formatted: Font: Italic, Complex Script Font: Italic

II. 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 that fall in different clusters. In real-world data, this assumption may not always hold~~ in each data~~, as illustrated in t~~. T~~he following example~~ illustrates this situation.~~, I~~I~~n this example the data includes ~~2 ~~two classes (circles and diamonds). ~~Suppose that ~~If we cluster ~~this~~the data into two clusters, then ~~we will get that~~ the left cluster will include two types of classes and the right one ~~is pure (i.e., ~~ will still have all the points from the same class~~)~~.

Formatted: Font color: Red, Strikethrough

Formatted: Strikethrough

Formatted: Font color: Red

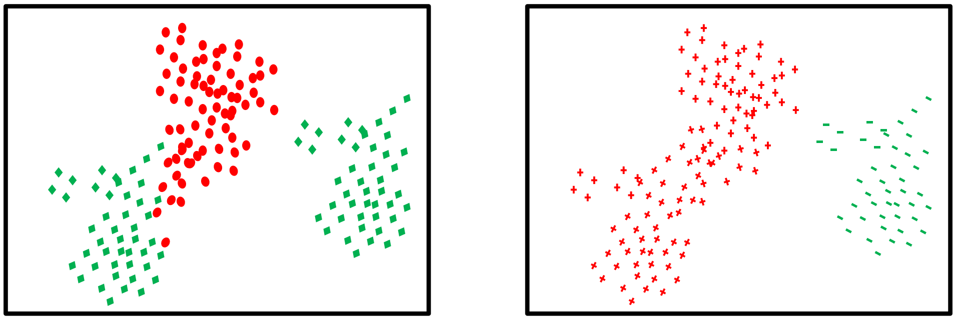
Formatted: Strikethrough

Formatted: Font color: Red

Formatted: Strikethrough

Formatted: Font color: Red

Formatted: Font color: Red, Strikethrough



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

Formatted: Font color: Red

Formatted: Font color: Red

A. 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, D be a set~~–~~ of labeled observations ~~that ~~used as ~~a ~~training data, and A ~~is ~~a set of unlabeled data. First~~ly~~, the algorithm will construct E, where E is a dataset ~~of ~~combining~~–~~ D and A (i.e., $E = D \cup A$), then the algorithm runs the k-means clustering algorithm several times with different values of k (we refer it to *nmc* = number of clusters)~~–~~ and ~~constructing~~builds the clustering matrix $cMat$~~-~~. $cMat$~~–~~ is a matrix where the i^{th} row consists of the clustering results of the i^{th} object in E. See Table 1 for an example.

Formatted: Font color: Red, Strikethrough

Formatted: Font color: Red

Formatted: Font color: Red, Strikethrough

Formatted: Font color: Red

~~At the end~~The end result is that~~,~~ each $x_i \in E$ is transformed into a new sample $x_i^* \in cMat$ with categorical values. The dimension of the x_i^* is *k*. ~~We should emphasize~~Please note that one needs to ~~consider ~~take into account the 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 an example of 20 samples with *k*=11. We record the results from *k*=2 as with *k*=1 – all the samples ~~will be~~are placed in one cluster.

Table 1: EC space for 20 samples and number of cluster (nmc) of 11. ~~F~~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 result~~s~~ of assigning k-means for each sample into 3 clusters etc. ~~and so on.~~

Formatted: Font color: Red

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

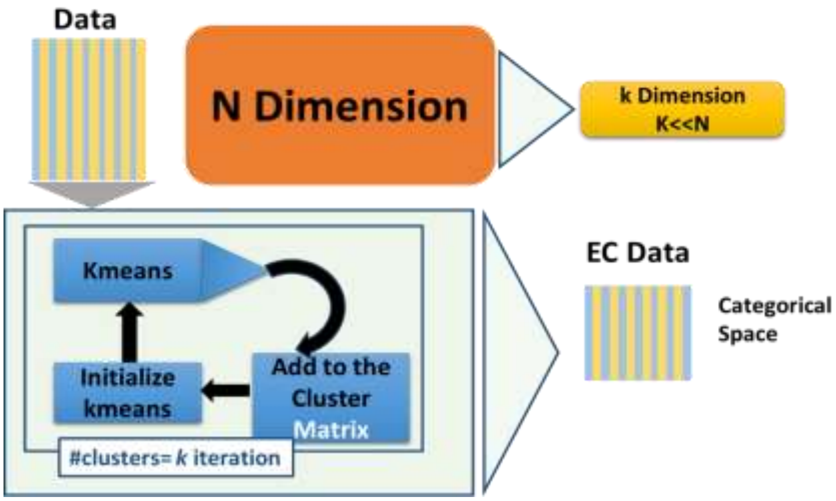


Figure 1*: The workflow for creating the EC categorical space based on the k-means clustering algorithm. The original data is the input to the workflow. The outcome is a new dataset named EC data in a categorical space with dimension k. the sign << indicates that k is -dramatically smaller -than the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

Formatted: Font color: Red

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

```
EC Transformation
Input:
E(l,N) : $x_1$ , $x_2$,…,$x_l$  Data consists of $l$ samples in N dimension(features)
k: number of clusters
~~c~~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 $x_i$ a cluster $c_0,c_1,..,c_{k-1}$
    (see Table 1 for an example of cMat)
```

Formatted: Font color: Red

B. Reduction of the EC sample

The new categorical data that results ~~of~~from applying the EC transformation (*Algorithm 1*)~~,~~ consists of l samples with k categorical features. As a results, t~~T~~he 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 sample dimension ~~could~~ can also be reduced ~~in terms 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~~ in Table 1, sample 11, sample 12 and sample 20 have the same categorical values. The vector space that represents those 3 points is g=(c0, c2, c2, c2, c4, c5, c6, c5, c5, c4), additionally, as a result, the points (samples) sample 1 and sample 18 have the same values and can then ~~can~~ be represented by and reduced to one point. ~~Therefore, t~~The new EC samples then become ~~are~~ redundant and can be represented by ~~based on the representors~~ g_i.
We have iterated all over the points in the EC data and keep the representor for each group.
Note that, the set E contains labeled and unlabeled data, and 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 the~~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: here~~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~~ as in (2).
4. ~~All the objects are not~~Not all the objects are labeled: in this ~~case~~case, the group will be an unclassified group.

Formatted: Font color: Red

Formatted: Font color: Red, Strikethrough

Formatted: Font color: Red

Comment [A14]: Unclear in the context – do you mean 'that'

Comment [A15]: What do you mean here please?

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

$$purity(g_i) = \sum_{j=1}^{\#classes} p_j^2$$

where g_i denotes group i that was represented by vector g_i in the matrix G, $\#classes$ denotes the number of the members, g_i, and p_j denote~~s~~ the probability of class j in group i. As can be seen, $purity(g_i)$ equals 1 when the group is pure and $\frac{1}{\#classes}$ for the lowest purity, that will decrease as the number of the classes increases.

III. 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~~goal- of the classifier is to generate a unique ~~the~~ EC ~~unique~~ sample~~s~~ from the generated EC samples, which ~~actually~~ is the representative set of EC samples. Next, ~~then~~ ~~we need to check~~ for each ~~represented~~ EC sample, we need to check the distribution of the labels in its original group.

Comment [A16]: Not entirely clear. Please clarify.

Algorithm 2 : Our new approach for classification-based EC is to ~~G~~grouping~~ing~~ the EC~~-~~-based Classifier ~~named Grb~~ (ClassifierEC)~~ is our new approach for classification based EC~~.

```
Grouping based classifier
Input:
$cMat$ a matrix with the ensemble clustering results.
E($l$,N) : $x_1$ , $x_2$, ...,$x_l$  Data consists of $l$ samples in N dimension( features)
$k$: number of clusters
C~~c~~reate empty matrix $cMat$ with $l$ rows (number of samples) and $k$ columns.
Algorithm:
1. Create the $groups$ based on the EC results.
2. For each $group_i$:
    2.1. Repeat until stopping criteria satisfies:
        2.1.1. Select labeled representor $g_i$.
        2.1.2. Assign the label of $g_i$ to all the unlabeled $group_i$ members.
    2.2. Classify all the unlabeled $group_i$ members by the majority class that they have.
    2.3. Calculate the $purity(group_i)$
    2.4. The accuracy for each unlabeled member will be the same as for the group purity.
3. Return the labeled dataset.
```

IV. Experiments on Numerical Datasets

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

V. DATASETS

The data ~~is consisting~~consists of microRNA precursor sequences, and ~~where~~ each sequence ~~consistsconsist of~~is made up of 4 nucleotide letters {A,U,C,G,}. T~~and t~~he length of each precursor sequence is about 70 nucleotides. The source of this~~such~~ data is miRbase[15]. Part ~~We have used part~~ of the data we have used has ~~that was used~~ was from other~~ in~~ different studies[16,17], including ~~and also used in~~ our ~~previouseprevious study~~ [13].

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

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

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

Table 2: The table shows a list of clades used in the study~~,~~. T~~t~~he first column represents the name of the clade, the second column the number of pre-cursors available on miRBase, and the third column the~~ 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: Ten datasets. The first column shows~~is~~ the name of the first clade positive data, and the second column ~~is~~ the second clade negative data.

Positive Data	Negative Data
Aves	Embryophyta
Cercopithecidae	Malvaceae

Formatted: Font color: Red, Strikethrough

Formatted: Font color: Red

Formatted: Strikethrough

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red, Strikethrough

Formatted: Font color: Red

Formatted: Font color: Red, Strikethrough

Formatted: Font color: Red

Formatted: Font color: Red, Strikethrough

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Formatted: Font color: Red

Embryophyta	Laurasiatheria
Fabaceae	Nematoda
Hexapoda	Aves
Laurasiatheria	brassicaceae
Malvaceae	Fabaceae
brassicaceae	Hexapoda
hominidae	Cercopithecidae
Monocotyledons	homoSapiens

VI. Reduction of the EC Sample

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

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

Table 4: 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~~ shows~~presents~~ the distribution of the group size for *k*=30 and *k*=50, and clearly indicates ~~. It is clear~~ that as ~~the~~ *k* ~~is~~ increases,~~ing~~ the number of groups with size 1 ~~is~~ also increases~~ing~~.~~—~~ ~~One expect~~The expectation is that ~~to get~~ the number of groups of size of 1 should~~to~~ be the same as the number of the original number of samples as we increase~~ing~~ the value of k. In other words, each sample will be hosted in one cluster.~~—~~ This actually raises a scientific question:~~,~~ what is the optimal value of *k* that will yield in improving the performance of the classifier, or more specifically, capture~~ing~~ the nature of the data in terms of clusters.

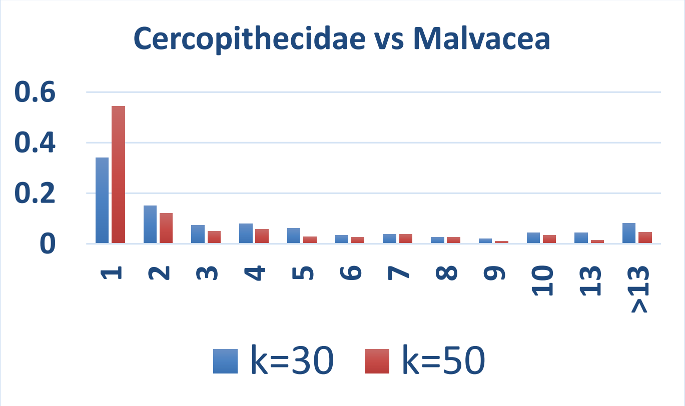


Figure 2:Distributaion of the groups samples (points) size comparing nmc=30 and nmc=50.

A. 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~~performed 100 iterations with equal sample size, and then calculated the mean of each performance measurements described below.

Comment [A17]: Please clarify

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 formula~~tion~~s (~~with~~ TP: ~~t~~True ~~p~~Positive, FP: ~~f~~False ~~p~~Positive, TN: ~~t~~True ~~n~~Negative, and FN ~~referring to f~~False ~~n~~Negative classifications):

$$Sensitivity = \frac{TP}{TP + FN} \; (SE, \text{recall})$$

$$Specificity = \frac{TN}{TN + FP} \; (SP)$$

$$Sensitivity = \frac{TP + TN}{TP + FN + TN + FP} \; (ACC)$$

B. Results

We also ~~have~~ conducted a ~~comparison~~ study ~~for the~~comparing 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 show~~ing~~ that the performance of the suggested classifier GrbClassifierEC~~–~~ was superior.

Figure 3 shows the performance of different classifiers at~~over~~ different levels of training percentage of the data. The results of EC ~~are~~ refer~~ring~~ to our own GrbClassifierEC classifier. We see that the performance is not significantly~~dramatically~~ influence~~s~~d by the size of the training part for the other classifiers while it does increase~~ing~~ significantly ~~dramatically~~ for the GrbClassifierEC classifier, at the 39% level. ~~Moreover~~In addition~~,~~ ~~it could reach a very high~~ performance can be improved significantly ~~as the perchance of~~if the training part is increased, ~~ing which is actually~~as a function of the value of k in the EC transformation.

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

Table 5 and Table 6 show~~s~~ the ~~comparison~~ results of a comparison 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).

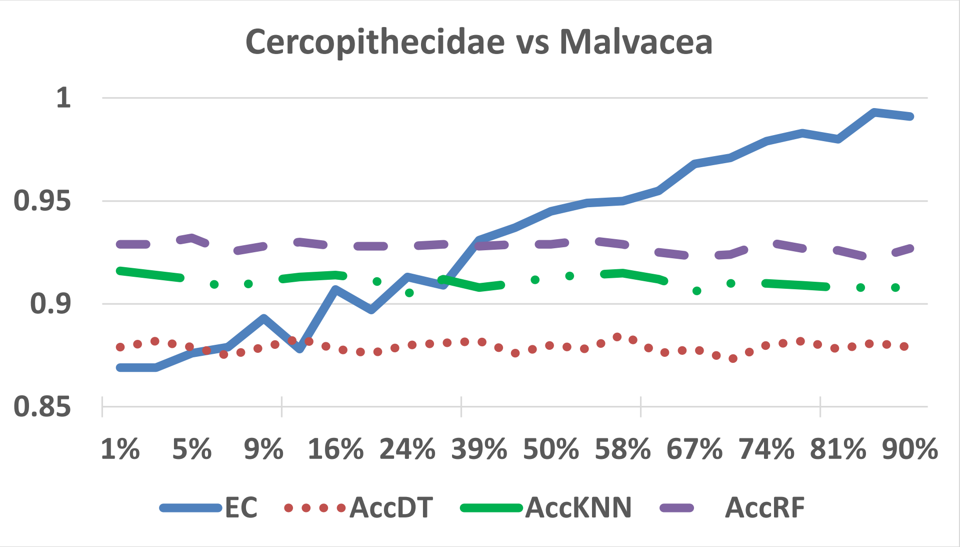


Figure 3: The accuracy of the classifiers over different level of sample training size.

Table 5 presents results with a *k* value of 49, while ~~t~~Table 6 presents 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 use~~ing~~ 80% of the data for training. The EC classifier ~~is~~ outperforms~~ing~~ the standard~~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 a k value 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%**

The results in Table 6 ~~demonstrate~~show 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 outperforms DTT and EC-RF and KNN with 5%, 3% and 1% respectively, while RF outperforms it with 2%. More interestingly, that ration of the reduction is an indication about the data redundan~~t~~cy and the similarity of the original data points ~~of the data~~.

Comment [A18]: Do you mean: …one *ca* reduces the size of the data to 39%..

Comment [A19]: Do you mean ‘at’ or ‘by’?

Comment [A20]: Ratio?

Table 6: GrbClassifierEC:~~–~~ -EC classifier results with a k value 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 positive value ~~of positive means~~indicates that the EC classifier is better than the other corresponding classifiers. EC-RF is a random forest applied on the EC data, RF is a random forest applied on the original data. DTT is a decision tree~~ss~~ 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%

VII. Conclusion

In this ~~work~~paper we ~~have ~~demonstrated the advantage of the EC approach in reducing the feature space and also in reducing the data size. In a~~A~~ddition~~ally~~, we ~~have ~~proposed ~~a new classifier approach named~~ using the new 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 achieve a ~~get ~~similar performance level, or even outperform ~~to ~~regular classifiers applied on the original data.~~ or even in some cases outperform them~~. However, to achieve th~~o~~ese results ~~are obtained in a pay off in~~the computation times that the ES transformation algorithm requires, increase.

The main assumption was that~~,~~ points within the same cluster share common traits more than points within different clusters. Thus it may be more beneficial to~~,~~ represent~~ing the~~ objects based on the clustering space rather ~~it may be better~~ than the geometric space.

The approach suggested here is very useful for the field of big data that allows~~ed~~ a~~to~~ reduction~~e~~ ~~the~~of the data to~~ a~~ representative data, by taking into account~~considering its~~ the EC data. For~~As a~~ future ~~work~~research we will need to suggest an~~d~~ algorithm that would pick the optimal value of k that and ~~would~~ yield ~~in ~~improved~~ing the~~ performance ~~under the constrains of~~while ~~–~~reducing the size of the data considerably~~dramatically~~.

Our algorithm~~, however, is general and~~ can be integrated with many other algorithms. In this research, we use only the k-means clustering algorithm with different k values. In ~~the ~~future ~~work~~research, we~~there~~

Comment [A21]: Tree?

Comment [A22]: Can you help clarify this please?

Formatted: Justified, Indent: First line: 0.36 cm

~~are~~ propose several directions: (1) checking the effect of the clustering algorithm to build an ensemble clustering space. (2) ~~how to detect~~finding poor clustering results based on the training data,~~.~~ (3) reducing the volume of the data by combining~~e~~ similar points 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.