Text Mining: High-Dimensional datasets clustering

Ricardo Navares

Abstract

Clustering is an unsupervised technique for grouping a set of objects in such way that objects in the same group are more similar in some sense or another to each other than to those in other groups. There are several decisions to be taken before tackling the problem such as the number of groups to use, the measures to use in order to decide which element belongs to which group... These questions will be addressed showing the advantages and disadvantages of each configuration.

Introduction

Clustering is the task to classify objects without previous knowledge of the class they belong as oppossed to classification. A wrapper to CLUTO[1] software is used to access this tool for high-dimensional datasets clustering on a collection of documents from Reuters-2178

The first part serves as a review of the theoretical concepts needed to understand the algorithms. Concretely, agglomerative clustering and K-means will be covered. The main difference between both techniques is that K-means initialize the grouping from a predefined number of centroids while in the agglomerative case, each element represents a cluster and iteratively the are grouped to form bigger clusters. Clustering consists on either moving the centroids or sequentially forming small clusters until the space of elements are combined in a predefined number of groups. Different measures and functions used to build the clusters will be covered in the case of the agglomerative algorithm, for K-means these functions define cecntroids relocation.

Finally, we will compare all possible configurations based on the quality measures and entropy and purity of the clusters. For this specific example

[1] http://glaros.dtc.umn.edu/gkhome/views/cluto/

September 8, 2017

Collection	Source	num. Documents	num. Terms	num. Classes
re0	Reuters-21578	1504	2886	13
re1	Reuters-21578	1657	3758	25

Table 1: *re0* and *re1* specifications

an optimal solution will be provided to help the reader to understand the process step by step.

Documents

The documents belong to the Reuters-21578 collection, indexed and categorized by Reuters. This set is distributed in 22 files, each of them containing 1000 documents in SGML format. Please notice this description is based on version 1.0 (2004) [1] and its constent might be updated. Concretely, collections *re0* and *re1* was used as shown in Table 1

theoretical background

Document clustering consists on applying analysis techniques to documents in order to organize, categorize and extract information from them. The most common algorithms lay in two main types, hierarchical and based on K-means and their variants [2].

Hierarchical algorithms generate a dendrogram where each intermediate levels can be interpreted as the combination of the clusters of its inferior level. There are two types of hierarchical cluster, agglomerative and divisive. We will focus on the agglomerative kind. The task starts with one cluster per individual element or class and, sequentially at each step combines pairs according to their similarity.

As opposed to hierarchical clustering the are several partition-based techniques. Since these ones are numerous we will focus on basic K-means algorithm which is the most commonly used. Given a predefined number of K clusters, the centroids are defined, these can be interpret as the center of each group based on a distance metric. At each step the ceentroids are updated trying to minimize the distance to each element of the cluster.

Similarity measures

A similarity measure is a function to quantify how similar two elements are. It can be interpret as a measure inversed to distance, it is, longer distance lower similarity and viceversa. One of the most common similarity measures used [3] is provided by the cosine function,

$$cos(d_i, d_j) = \frac{d_i^t \cdot d_j}{\|d_i\| \, \|d_j\|} \tag{1}$$

Given this function is defined for vectors and it is intended to classify documents, unary vectors can be used to simplify (1) to $cos(d_i, d_j) = d_i^t \cdot d_j$. It is clearly interpretable and it is related to the orientation of the vector. Vectors with identical orientation (0°), will have a cosine equal to 1. Similarly, two orthogonal vectors (90°) have a cosine equal to 0 and consequently the will not be similar. Same applies to documents.

The second most common measure proposed by [3] is the euclidean distance but it is defined for criterion functions based on graphs and does not apply in this case. As an alternative we can use the correlation defined by (2)

$$corr(d_i, d_j) = \frac{(d_i - \overline{d_i})(d_j - \overline{d_j})}{\|d_i - \overline{d_i}\| \, \|d_j - \overline{d_j}\|} \tag{2}$$

Where $\overline{d_i}$ and $\overline{d_j}$ are the means. By definition, correlation takes values in the interval $[-1, 1]$. Being 1 a perfect correlation, ergo identical documents and -1 representing completely different documents.

Criterion functions

Given a cluster contains several elements or objects, in practice there are many candidates on which the similarity measures can be applied. As this number can be extremely high, there is a need to apply a criteria in order to select a reduce set on which the functions will be applied. There are 4 types of criterion functions [3]: internal, external, hybrid and graph-based. We will only cover two of them, an internal $\mathcal{I}_2$ and a hybrid one $\mathcal{H}_1$ based on the results obtained in Zhao *et al.* [3] which show for these two metrics better entropies and purities.

An internal function only takes into account those elements which belong to a cluster, ignoring the elements that were assign to other clusters. Concretely, it optimize the function define for the set of docuuments that belong to the cluster,

$$\text{maximize } \mathcal{I}_2 = \sum_{r=1}^{k} \sum_{d_i \in S_r} cos(d_i, C_r) \tag{3}$$

being k the number of clusters, S_i the subsety of elements (documents) of cluster i and C_i the centroid of the cluster. An hybrid function combine those internal and external functions. It is evident that an external function, as opposed to internal, takes only into consideration thposition of other clusters. An hybrid function is defined as follows,

$$\text{maximize } \mathcal{H}_1 = \frac{\mathcal{I}_1}{\mathcal{E}_1} \tag{4}$$

where the externl function is defined by the similarity cosine functions between the centroid of cluster C_r and the vector of remaining centroid C weighting by the nuumber of elements n in the cluster.

$$\mathcal{E}_1 = \sum_{r=1}^{k} n_r \cdot cos(C_r, C) \tag{5}$$

and the internal function is defined by the sum of similarity means of the documents of each cluster weighted by the size of the cluster.

$$\mathcal{I}_1 = \sum_{r=1}^{k} n_r \left(\frac{1}{n_r^2} \sum_{d_i, d_j \in S_r} cos(d_i, d_j) \right) \tag{6}$$

Algorithm evaluation

In order to evaluate the quality of an algorithm we can use the entropy which measures how the elements (documents) are distributed in each cluster,

$$E(S_r) = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \tag{7}$$

where q is the class which contains the document and n_r^i the number of documents of class i. The entropy of the solution is defined by the sum of entropies of the clusters weighted by the size of the cluster,

$$E = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r) \tag{8}$$

It can be clearly seen that a cluster with only one class q of element (document) leads in (7) to the condition $n_r^i = n_r$ and, consequently, the logarithm is equal to 0, thus a perfect classification. As a second measure, purity represents the most representative class in each cluster weighted by the cluster size,

$$P(S_r) = \frac{1}{n_r} \max_i (n_r^i) \tag{9}$$

being i the class. The total purity is defined by,

$$P = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r) \tag{10}$$

Again, a perfect clustering condition implies that each cluster contains only one class so $n_r = \max_i(n_r^i)$ an the total purity will be equal to the number of clusters k.

Material and methods

The experiments are taylored using CLUTO[2] through command line *vcluster*. This tool provides an interface to the implementation of the methods described in the previous sections. Several configurations will be tests in order to compare the metrics and evaluate the idoneity of each solution.

Execution

The wrapper *main.sh* to CLUTO was written in the scriptting language BASH [3] which asks the user how many clusters are wanted to be executed. The results are provided in text files for each execution. The following files are needed:

```
../
    - vcluster
    - main.sh
    - re0.mat
    - re0.mat.rclass
    - re1.mat
    - re1.mat.rclass
```

[2] http://glaros.dtc.umn.edu/gkhome/views/cluto/

[3] https://www.gnu.org/software/bash/

```
echo -n "Enter number of clusters for re0 [ENTER]: "
read nclusters
algorithm=("direct" "agglo")
similarity=("cos" "corr")
criteria=("i2" "h1")

echo -ne "Calculating re0"
for i in ${algorithm[@]}; do
  for j in ${similarity[@]}; do
    for k in ${criteria[@]}; do
        cmd="./vcluster -rclassfile=re0.mat.rclass -clmethod=$i
              -sim=$j -crfun=$k re0.mat $nclusters";
        eval $cmd>>resultados_re0.txt;
        echo -ne ".";
    done
  done
done
echo "Finish"

echo -n "Enter number of clusters for re1 [ENTER]: "
read nclusters

echo -ne "Calculating re1"
for i in ${algorithm[@]}; do
  for j in ${similarity[@]}; do
    for k in ${criteria[@]}; do
        cmd="./vcluster -rclassfile=re1.mat.rclass -clmethod=$i
              -sim=$j -crfun=$k re1.mat $nclusters";
        eval $cmd>>resultados_re1.txt;
        echo -ne ".";
    done
  done
done

echo "Finish"

exit 0;
```

Command vcluster

It is required to execute CLUTO via command line with the following parameters:

- **-rclassfile**: File *.mat.rclass* which contains the classes to which the docuemnts belong in order to identify the clusters.

- **-clmethod**: Algorithm to be used. *"direct"* for K-means and *"agglo"* agglomerative

- **-sim**: Similarity measure. *"cos"* cosine and *"corr"* correlation

- **-crfun**: Criterion function. *"i2"* for $\mathcal{I}_2$ and *"h1"* for $\mathcal{H}_1$

- **fichero.mat**: File with the data
- **n**: Number of clusters (K)

```
./vcluster -rclassfile=  -clmethod=  -sim= -crfun= fichero.mat n
```

Number of clusters

The scripts results show the entropy for each configuration. At first sight, it makes sense to define a number of clusters equal to the number of classes of each collection as shown in Table 1. It is 13 for *re0* and 25 for *re1*, as in the execution of the following part,

```

echo "Calculating optimal number of clusters for re0:"
clusters=("5" "7" "8" "9" "10" "11" "12" "13")

for i in ${clusters[@]}; do
  cmd="./vcluster re0.mat $i |
        grep -o 'Entropy: 0.[0-9][0-9][0-9]' |
        sed 's/\Entropy: //g'"
    entropy=$(eval $cmd)
    echo "Entropy for $i clusters: $entropy"
done

echo "Calculating optimal number of clusters for re1:"
clusters=("5" "10" "15" "20" "21" "22" "23" "24" "25")

for i in ${clusters[@]}; do
    cmd="./vcluster re1.mat $i |
        grep -o 'Entropy: 0.[0-9][0-9][0-9]' |
        sed 's/\Entropy: //g'"
    entropy=$(eval $cmd)
    echo "Entropy for $i clusters: $entropy"
done
```

provides,

```
Calculating optimal number of clusters for re0:
Entropy for 5 clusters: 0.518
Entropy for 7 clusters: 0.448
Entropy for 8 clusters: 0.445
Entropy for 9 clusters: 0.412
Entropy for 10 clusters: 0.403
Entropy for 11 clusters: 0.397
Entropy for 12 clusters: 0.396
Entropy for 13 clusters: 0.395
Calculating optimal number of clusters for re1:
Entropy for 5 clusters: 0.476
Entropy for 10 clusters: 0.406
Entropy for 15 clusters: 0.351
Entropy for 20 clusters: 0.314
Entropy for 21 clusters: 0.309
Entropy for 22 clusters: 0.309
Entropy for 23 clusters: 0.305
Entropy for 24 clusters: 0.299
Entropy for 25 clusters: 0.298
```

It can be clearly see that the entropy decreases as the number of clusters increase, although its value is problem-specific and depending on what the user wants it can be set to a different number depending on the needs. In order to compare the metrics and to save execution time, henceforth we will set $K = 5$ which also makes it easier to show the results in tables.

Results and discussion

Algorithm comparison

To compare the algorithms, as a first step of the analysis we are using the default values with *CLMethod=XXX* indicating the algorithm *XXX* which will be ”agglo” or ”direct”.

```
./vcluster -rclassfile=re0.mat.rclass -clmethod="direct" re0.mat 5
./vcluster -rclassfile=re0.mat.rclass -clmethod="agglo" re0.mat 5
./vcluster -rclassfile=re1.mat.rclass -clmethod="direct" re1.mat 5
./vcluster -rclassfile=re1.mat.rclass -clmethod="agglo" re1.mat 5

Options ---------------------------------------------------------------------
  CLMethod=XXX, CRfun=I2, SimFun=Cosine, #Clusters: 5
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10
```

Table 2 shows the entropies and the purities of each algorithm applied on each collection of documents. It is shown that the K-menas obtains lower entropies for both cases with also higher purities.

		Entropy	Purity
K-means	re0	0.516	0.471
	re1	0.479	0.556
Agglomerative	re0	0.623	0.465
	re1	0.765	0.274

Table 2: Algorithm comparison: clustering quality

Algorithm	cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty
K-mean	0	117	0.458	0.128	0.033	0.008	0.16	0.906
	1	175	0.189	0.072	0.034	0.013	0.357	0.72
	2	408	0.094	0.03	0.031	0.012	0.932	0.145
	3	252	0.082	0.025	0.032	0.015	0.32	0.639
	4	552	0.06	0.02	0.026	0.012	0.423	0.466
Agglomerative	0	1	1	0	0.013	0	0	1
	1	3	0.427	0.044	0.021	0.004	0.428	0.333
	2	485	0.061	0.021	0.028	0.013	0.436	0.458
	3	80	0.164	0.053	0.026	0.01	0.026	0.988
	4	935	0.063	0.023	0.028	0.013	0.772	0.425

Table 3: Algorithm comparison::Internal and external metrics for re0

The reason is shown in Tables 3 and 4 where cluster C_4 is way bigger compared to the rest, contianing many objects of different classes. As a result it gets high entropy and low purity,

Algorithm	cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty
K-means	0	194	0.10	0.04	0.02	0.01	0.49	0.48
	1	225	0.09	0.03	0.02	0.01	0.52	0.44
	2	402	0.06	0.02	0.02	0.01	0.35	0.74
	3	480	0.05	0.02	0.02	0.01	0.47	0.66
	4	356	0.04	0.01	0.01	0.01	0.61	0.33
Agglomerative	0	8	0.28	0.07	0.01	0.00	0.41	0.38
	1	8	0.21	0.04	0.01	0.00	0.23	0.75
	2	3	0.41	0.04	0.01	0.01	0.20	0.67
	3	176	0.06	0.02	0.01	0.01	0.40	0.58
	4	1462	0.03	0.01	0.01	0.01	0.82	0.23

Table 4: Algorithm comparison::Internal and external metrics for re1

For *re0* in Table Appendix A.1 it can be seen that clusters C_0 and C_1 contain very few classes for agglomertative clustering. Situation extended to C_2 for *re1* collection in Table Appendix A.2. High purity is achieved in these cases but clusters with very few clases are not convenient in this example as their entropy and the external metrics are low in similar proportion.

Comparison similarity functions

Shell command to compare similarity functions and the default options needed are shown below. Notice that *SimFun=XXX* specifices function *XXX* which can be cosiene (”Cosine”) or correlation (”CorrCoef”).

		Entropy	Purity
Cosine	re0	0.518	0.513
	re1	0.476	0.543
Correlation	re0	0.498	0.580
	re1	0.560	0.512

Table 5: Comparison similarity functions: clustering quality

```
./vcluster -rclassfile=re0.mat.rclass -sim="cos" re0.mat 5
./vcluster -rclassfile=re0.mat.rclass -sim="corr" re0.mat 5
./vcluster -rclassfile=re1.mat.rclass -sim="cos" re1.mat 5
./vcluster -rclassfile=re1.mat.rclass -sim="corr" re1.mat 5

Options -------------------------------------------------------------------------
  CLMethod=RB, CRfun=I2, SimFun=XXX, #Clusters: 5
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10
```

We can see in Table 5 which function fits to each set. It looks like the cosine function obtains better results both entropy and purity for *re0* collection while te correlation function fits better for *re0*. Tables 6 and 7 show that the correlation function along with the external measures are bigger when the cosine function is used, telling a bigger cluster differentiation.

Tables Appendix A.3 and Appendix A.4 show the clusters per class. Special attention deserves Table Appendix A.3 for *re0* collection. Using the cosine function, cluster number 2 contains similar amount of elements for each class, consequently its purity decreases which can be supplemented with Table 5 that it shows a more equitative distribution. On the other hand Table Appendix A.4 shows a better distribution of the classes for *re1*.

Function	cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty
Cosine	0	127	0.41	0.14	0.03	0.01	0.22	0.86
	1	185	0.18	0.07	0.04	0.01	0.31	0.74
	2	347	0.10	0.03	0.03	0.01	0.94	0.14
	3	377	0.07	0.02	0.03	0.01	0.31	0.57
	4	468	0.06	0.02	0.03	0.01	0.54	0.56
Correlation	0	139	0.49	0.14	0.10	0.03	0.24	0.85
	1	286	0.26	0.07	0.08	0.03	0.61	0.45
	2	572	0.23	0.07	0.07	0.03	0.79	0.31
	3	258	0.18	0.05	0.05	0.03	0.17	0.90
	4	249	0.18	0.07	0.07	0.03	0.20	0.87

Table 6: Comparison similarity functions: External and internal measures for re0

Function	cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty
Cosine	0	203	0.10	0.04	0.02	0.01	0.49	0.47
	1	314	0.07	0.03	0.02	0.01	0.63	0.32
	2	368	0.06	0.02	0.02	0.01	0.26	0.81
	3	393	0.06	0.02	0.02	0.01	0.37	0.73
	4	379	0.04	0.01	0.01	0.01	0.66	0.31
Correlation	0	447	0.16	0.06	0.04	0.02	0.67	0.42
	1	459	0.13	0.05	0.04	0.02	0.46	0.66
	2	213	0.12	0.06	0.04	0.02	0.62	0.39
	3	263	0.08	0.03	0.04	0.02	0.45	0.66
	4	275	0.07	0.03	0.03	0.02	0.60	0.38

Table 7: Comparison similarity functions: External and internal measures for re1

		Entropy	Purity
I2	re0	0.518	0.513
	re1	0.476	0.543
H1	re0	0.515	0.508
	re1	0.504	0.551

Table 8: Comparison criterion functions: clustering quality

Criterion functions comparison

Line commands are defined as follows. Notice that *CRfun=XX* specifies function *XX*, being either hybrid (”H1”) or internal (”I2”)

```
./vcluster -rclassfile=re0.mat.rclass -crfun="i2" re0.mat 5
./vcluster -rclassfile=re0.mat.rclass -crfun="i2"" re0.mat 5
./vcluster -rclassfile=re1.mat.rclass -crfun="i2" re1.mat 5
./vcluster -rclassfile=re1.mat.rclass -crfun="i2" re1.mat 5

Options ----------------------------------------------------------------------
  CLMethod=RB, CRfun=XX, SimFun=Cosine, #Clusters: 5
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H1, NTrials=10, NIter=10
```

Table 8 shows better results for *re0* for the hybrid function as it obtains higher entropy and purity, this does not happen for the collection *re1*. Internal and external metrics are shown in Tables 9 and 10. It is obtained in $\mathcal{H}_1$ similart internal an external metrics values which defines the convenience of this function. Table Appendix A.5 for *re0* clearly shows the term ”mone” with high representation on each cluster so the algorithms does not manage to classify it propoerly. Besides, there are significative differences between both functions when categorizing the classes of this collection. In *re1* the results are similar for both metrics as shown in Table Appendix A.6.

Function	cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty
I2	0	127	0.41	0.14	0.03	0.01	0.22	0.86
	1	185	0.18	0.07	0.04	0.01	0.31	0.74
	2	347	0.10	0.03	0.03	0.01	0.94	0.14
	3	377	0.07	0.02	0.03	0.01	0.31	0.57
	4	468	0.06	0.02	0.03	0.01	0.54	0.56
H1	0	108	0.51	0.10	0.03	0.01	0.09	0.95
	1	157	0.22	0.07	0.04	0.01	0.31	0.74
	2	389	0.10	0.03	0.03	0.01	0.93	0.17
	3	393	0.07	0.02	0.03	0.01	0.31	0.55
	4	457	0.06	0.02	0.03	0.01	0.51	0.57

Table 9: Comparison criterion functions: IOnternal and external metrics for re0

Function	cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty
I2	0	203	0.10	0.04	0.02	0.01	0.49	0.47
	1	314	0.07	0.03	0.02	0.01	0.63	0.32
	2	368	0.06	0.02	0.02	0.01	0.26	0.81
	3	393	0.06	0.02	0.02	0.01	0.37	0.73
	4	379	0.04	0.01	0.01	0.01	0.66	0.31
H1	0	146	0.14	0.04	0.02	0.01	0.32	0.60
	1	200	0.10	0.04	0.02	0.01	0.44	0.48
	2	494	0.05	0.02	0.02	0.01	0.45	0.60
	3	474	0.05	0.02	0.02	0.01	0.47	0.66
	4	343	0.03	0.01	0.02	0.01	0.74	0.36

Table 10: Comparison criterion functions: IOnternal and external metrics for re1

Putting all together

Table 11 shows all the combinations of the selecter parameters. Agglomerative clustering with the correlation function and $\mathcal{H}_1$ obtains lower entropies and higher purities. As a general approach, it is more convenient the correlation function as a similarity measure either for both algorithms. On hte other hand, the results for *re1* are the contrary, being the cosine function with $\mathcal{I}_2$ the best configuration of metrics as shown in Table 12.

Conclusions

We have checked diverse cluster configurations and metrics to evaluate the performance of the algorithms over two collection of documents from Reuters. Give the two different problem characteristics we saw that an optimal configuration for one problem might not apply to other. Additionally

		Entropy		Purity	
		K-means	Agglomerative	K-means	Agglomerative
Cosino	I2	0.516	0.546	0.471	0.465
	H1	0.512	0.569	0.465	0.527
Correlation	I2	0.491	0.497	0.548	0.476
	H1	0.498	**0.475**	0.529	**0.559**

Table 11: Comparison re0 with K=5

		Entropy		Purity	
		K-means	Agglomerative	K-means	Agglomerative
Cosino	I2	**0.479**	0.564	**0.556**	0.482
	H1	0.489	0.571	0.551	0.461
Correlation	I2	0.548	0.571	0.513	0.499
	H1	0.633	0.565	0.414	0.492

Table 12: Comparison re1 with K=5

we exposed the need to research the influence of each parameter over each algorithm to obtain optimal results.

References

[1] David D. Lewis. *Reuters-21578 text categorization test collection.* `http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt`. 2004.

[2] Michael Steinbach, George Karypis, Vipin Kumar. Dept. of Computer Eng. University of Minnesota *A comparison of document clustering techniques.* 2000.

[3] Ying Zhao, George Karypis. Dept. of Computer Eng. University of Minnesota. *Criterion functions for document clustering. Experiments and analysis.* 2002.

class	K-means					Agglomerative				
cid	0	1	2	3	4	0	1	2	3	4
hous	0	0	15	0	1	0	0	0	0	16
mone	106	126	51	82	243	0	1	210	0	397
trad	5	1	54	2	257	0	1	222	1	95
rese	0	12	23	1	6	0	0	5	0	37
cpi	0	2	54	1	3	0	0	2	0	58
inte	4	29	4	161	21	1	1	32	79	106
gnp	0	1	59	4	16	0	0	5	0	75
reta	0	0	20	0	0	0	0	2	0	18
ipi	0	0	36	0	1	0	0	3	0	34
jobs	0	4	32	0	3	0	0	1	0	38
lei	0	0	11	0	0	0	0	2	0	9
bop	2	0	34	1	1	0	0	1	0	37
wpi	0	0	15	0	0	0	0	0	0	15

Table Appendix A.1: Algorithm comparison: Class classification for re0

class	K-means					Agglomerative				
cid	0	1	2	3	4	0	1	2	3	4
coco	42	0	0	6	0	0	0	0	0	48
grai	1	39	3	316	12	2	1	2	25	341
veg	0	45	4	10	1	0	0	0	1	59
whea	0	1	0	18	0	0	0	0	0	19
copp	1	0	0	0	36	0	0	0	0	37
coff	93	0	3	3	0	1	0	0	0	98
suga	1	99	0	5	1	3	0	0	1	102
ship	2	0	4	14	117	2	0	0	102	33
cott	1	0	0	16	0	0	0	0	0	17
carc	2	6	0	5	7	0	1	0	0	19
crud	0	1	297	0	32	0	0	0	33	297
nat	0	0	27	0	0	0	0	0	0	27
meal	0	2	0	13	0	0	0	0	1	14
alum	0	0	1	4	26	0	0	0	0	31
oils	0	10	0	33	7	0	0	1	6	43
gold	0	0	0	0	87	0	0	0	1	86
tin	11	0	1	1	5	0	0	0	5	13
live	6	13	0	20	3	0	6	0	0	36
iron	0	4	7	7	13	0	0	0	0	31
rubb	26	0	2	2	2	0	0	0	0	32
zinc	0	2	1	0	7	0	0	0	0	10
oran	6	0	1	6	0	0	0	0	0	13
pet	2	1	14	1	0	0	0	0	1	17
dlr	0	0	20	0	0	0	0	0	0	20
gas	0	2	17	0	0	0	0	0	0	19

Table Appendix A.2: Algorithm comparison: Class classification for re1

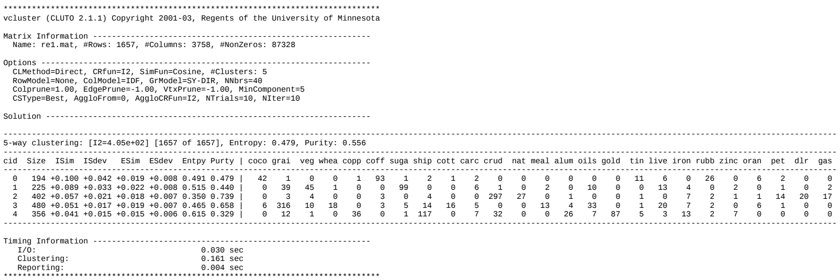
class	Cosine					Correlation				
cid	0	1	2	3	4	0	1	2	3	4
hous	1	0	14	0	1	0	2	14	0	0
mone	109	137	49	214	99	118	129	126	18	217
trad	6	0	49	3	261	10	56	9	233	11
rese	0	12	20	2	8	0	37	3	0	2
cpi	0	0	50	1	9	0	1	57	1	1
inte	9	32	7	156	15	2	24	175	1	17
gnp	0	0	26	1	53	2	2	73	2	1
reta	0	0	18	0	2	0	0	20	0	0
ipi	0	0	30	0	7	0	0	36	1	0
jobs	0	4	26	0	9	1	6	31	1	0
lei	0	0	10	0	1	0	0	11	0	0
bop	2	0	33	0	3	6	29	2	1	0
wpi	0	0	15	0	0	0	0	15	0	0

Table Appendix A.3: Similarity metrics: Class classification for re0

class	Cosine					Correlation				
cid	0	1	2	3	4	0	1	2	3	4
coco	48	0	0	0	0	8	2	36	2	0
grai	1	77	1	288	4	186	8	2	174	1
veg	0	46	7	6	1	19	38	0	2	1
whea	0	6	0	13	0	15	0	0	4	0
copp	0	1	0	0	36	8	13	3	1	12
coff	96	0	1	0	2	10	3	84	0	2
suga	2	99	0	5	0	80	0	13	11	2
ship	2	5	2	9	119	16	7	3	8	103
cott	1	1	0	14	1	4	1	3	8	1
carc	1	7	0	5	7	7	1	2	3	7
crud	1	1	297	1	30	1	302	0	0	27
nat	0	0	27	0	0	1	25	0	0	1
meal	0	10	0	5	0	13	1	0	1	0
alum	0	5	1	1	24	15	4	4	0	8
oils	0	25	0	18	7	25	7	0	11	7
gold	0	0	0	1	86	6	4	1	0	76
tin	9	2	0	0	7	6	1	5	0	6
live	7	13	0	19	3	6	5	7	22	2
iron	0	11	2	0	18	12	6	1	2	10
rubb	26	1	1	1	3	3	1	23	3	2
zinc	0	2	1	1	6	3	3	0	1	3
oran	7	0	0	6	0	0	0	4	9	0
pet	1	1	10	0	6	1	11	2	1	3
dlr	1	0	0	0	19	0	0	20	0	0
gas	0	1	18	0	0	2	16	0	0	1

Table Appendix A.4: Similarity metrics: Class classification for re1

Figure A.1: Solution re1 CLMethod=Direct, CRfun=I2, SimFun=Cosine.



```
********************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----------------------------------------------------------
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options ----------------------------------------------------------------------
  CLMethod=Direct, CRfun=I2, SimFun=Cosine, #Clusters: 5
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
5-way clustering: [I2=4.05e+02] [1657 of 1657], Entropy: 0.479, Purity: 0.556
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | coco grai  veg whea copp coff suga ship cott carc crud  nat meal alum oils gold  tin live iron rubb zinc oran  pet  dlr  gas
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
  0   194 +0.100 +0.042 +0.019 +0.008 0.491 0.479 |   42    1    0    0    1   93    1    2    1    2    0    0    0    0    0    0   11    6    0   26    0    6    2    0    0
  1   225 +0.089 +0.033 +0.022 +0.008 0.515 0.440 |    0   39   45    1    0    0   99    0    0    6    1    0    2    0   10    0    0   13    4    0    2    0    1    0    2
  2   402 +0.057 +0.021 +0.018 +0.007 0.350 0.739 |    0    3    4    0    0    3    0    4    0    0  297   27    0    1    0    0    1    0    7    2    1    1   14   20   17
  3   480 +0.051 +0.017 +0.019 +0.007 0.465 0.658 |    6  316   10   18    0    3    5   14   16    5    0    0   13    4   33    0    1   20    7    2    0    6    1    0    0
  4   356 +0.041 +0.015 +0.015 +0.006 0.615 0.329 |    0   12    1    0   36    0    1  117    0    7   32    0    0   26    7   87    5    3   13    2    7    0    0    0    0
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information -----------------------------------------------------------
   I/O:                                          0.030 sec
   Clustering:                                   0.161 sec
   Reporting:                                    0.004 sec
********************************************************************************
```

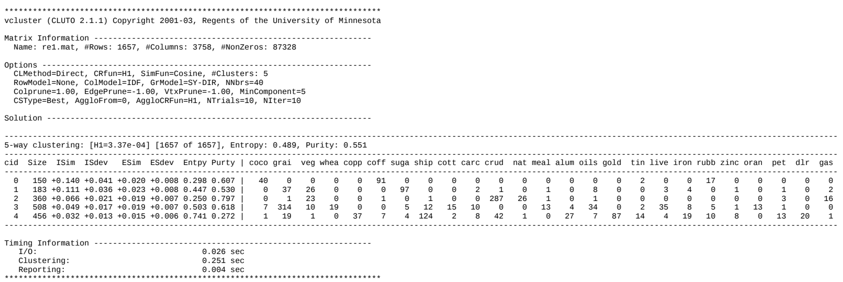
class	I2					H1				
cid	0	1	2	3	4	0	1	2	3	4
hous	1	0	14	0	1	0	0	15	0	1
mone	109	137	49	214	99	103	117	68	216	104
trad	6	0	49	3	261	2	0	54	3	260
rese	0	12	20	2	8	0	10	26	1	5
cpi	0	0	50	1	9	0	0	51	1	8
inte	9	32	7	156	15	2	26	6	170	15
gnp	0	0	26	1	53	0	0	31	2	47
reta	0	0	18	0	2	0	0	19	0	1
ipi	0	0	30	0	7	0	0	34	0	3
jobs	0	4	26	0	9	0	4	26	0	9
lei	0	0	10	0	1	0	0	10	0	1
bop	2	0	33	0	3	1	0	34	0	3
wpi	0	0	15	0	0	0	0	15	0	0

Table Appendix A.5: Comparison criterion functions: Class classification for re0

class	I2					H1				
cid	0	1	2	3	4	0	1	2	3	4
coco	48	0	0	0	0	38	0	0	9	1
grai	1	77	1	288	4	0	40	0	311	20
veg	0	46	7	6	1	0	40	7	11	2
whea	0	6	0	13	0	0	4	0	15	0
copp	0	1	0	0	36	1	0	27	0	9
coff	96	0	1	0	2	87	0	1	4	7
suga	2	99	0	5	0	1	97	0	6	2
ship	2	5	2	9	119	0	0	2	12	123
cott	1	1	0	14	1	0	0	1	14	2
carc	1	7	0	5	7	0	2	0	8	10
crud	1	1	297	1	30	0	1	295	0	34
nat	0	0	27	0	0	0	0	27	0	0
meal	0	10	0	5	0	0	1	0	14	0
alum	0	5	1	1	24	0	0	3	5	23
oils	0	25	0	18	7	0	9	0	33	8
gold	0	0	0	1	86	0	0	81	0	6
tin	9	2	0	0	7	2	0	0	3	13
live	7	13	0	19	3	0	5	1	15	21
iron	0	11	2	0	18	0	0	10	7	14
rubb	26	1	1	1	3	17	0	1	2	12
zinc	0	2	1	1	6	0	0	6	1	3
oran	7	0	0	6	0	0	0	0	4	9
pet	1	1	10	0	6	0	0	14	0	4
dlr	1	0	0	0	19	0	0	0	0	20
gas	0	1	18	0	0	0	1	18	0	0

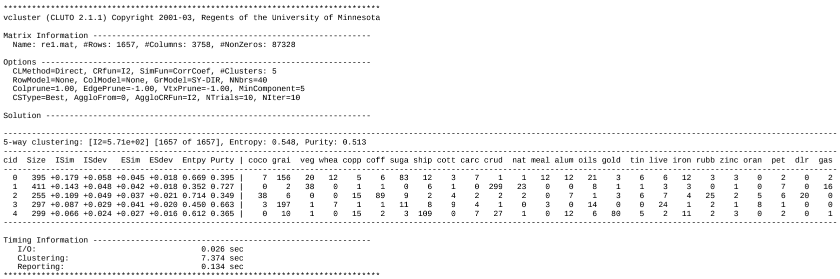
Table Appendix A.6: Comparison criterion functions: Class classification for re1

Figure A.2: Solution re1 CLMethod=Direct, CRfun=H1, SimFun=Cosine.



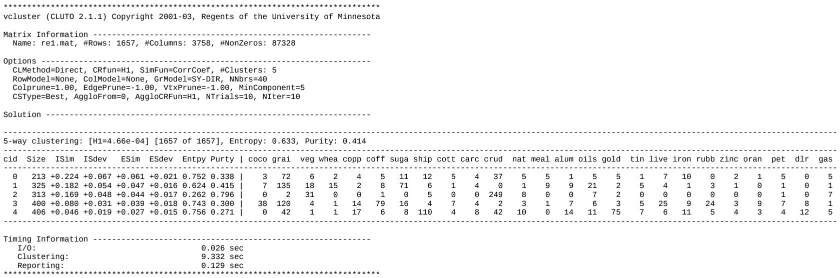
```
********************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----------------------------------------------------------
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options ----------------------------------------------------------------------
  CLMethod=Direct, CRfun=H1, SimFun=Cosine, #Clusters: 5
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H1, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
5-way clustering: [H1=3.37e-04] [1657 of 1657], Entropy: 0.489, Purity: 0.551
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | coco grai  veg whea copp coff suga ship cott carc crud  nat meal alum oils gold  tin live iron rubb zinc oran  pet  dlr  gas
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
  0   150 +0.140 +0.041 +0.020 +0.008 0.298 0.607 |   40    0    0    0    0   91    0    0    0    0    0    0    0    0    0    0    2    0    0   17    0    0    0    0    0
  1   183 +0.111 +0.036 +0.023 +0.008 0.447 0.530 |    0   37   26    0    0    0   97    0    0    2    1    0    1    0    8    0    0    3    4    0    1    0    1    0    2
  2   360 +0.066 +0.021 +0.019 +0.007 0.250 0.797 |    0    1   23    0    0    1    0    1    0    0  287   26    1    0    1    0    0    0    0    0    0    0    3    0   16
  3   508 +0.049 +0.017 +0.019 +0.007 0.503 0.618 |    7  314   10   19    0    0    5   12   15   10    0    0   13    4   34    0    2   35    8    5    1   13    1    0    0
  4   456 +0.032 +0.013 +0.015 +0.006 0.741 0.272 |    1   19    1    0   37    7    4  124    2    8   42    1    0   27    7   87   14    4   19   10    8    0   13   20    1
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information -----------------------------------------------------------
   I/O:                                     0.026 sec
   Clustering:                              0.251 sec
   Reporting:                               0.004 sec
********************************************************************************
```

Figure A.3: Solution re1 CLMethod=Direct, CRfun=I2, SimFun=CorrCoef.



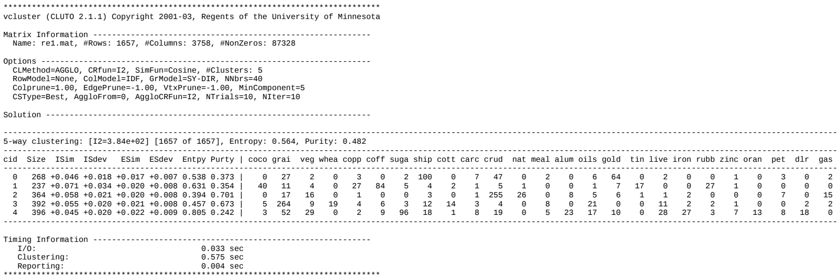
```
********************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----------------------------------------------------------
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options ----------------------------------------------------------------------
  CLMethod=Direct, CRfun=I2, SimFun=CorrCoef, #Clusters: 5
  RowModel=None, ColModel=None, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
5-way clustering: [I2=5.71e+02] [1657 of 1657], Entropy: 0.548, Purity: 0.513
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | coco grai  veg whea copp coff suga ship cott carc crud  nat meal alum oils gold  tin live iron rubb zinc oran  pet  dlr  gas
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
  0   395 +0.179 +0.058 +0.045 +0.018 0.669 0.395 |    7  156   20   12    5    6   83   12    3    7    1    1   12   12   21    3    6    6   12    3    3    0    2    0    2
  1   411 +0.143 +0.048 +0.042 +0.018 0.352 0.727 |    0    2   38    0    1    1    0    6    1    0  299   23    0    0    8    1    1    3    3    0    1    0    7    0   16
  2   255 +0.109 +0.049 +0.037 +0.021 0.714 0.349 |   38    6    0    0   15   89    9    2    4    2    2    2    0    7    1    3    6    7    4   25    2    5    6   20    0
  3   297 +0.087 +0.029 +0.041 +0.020 0.450 0.663 |    3  197    1    7    1    1   11    8    9    4    1    0    3    0   14    0    0   24    1    2    1    8    1    0    0
  4   299 +0.066 +0.024 +0.027 +0.016 0.612 0.365 |    0   10    1    0   15    2    3  109    0    7   27    1    0   12    6   80    5    2   11    2    3    0    2    0    1
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information -----------------------------------------------------------
   I/O:                                   0.026 sec
   Clustering:                            7.374 sec
   Reporting:                             0.134 sec
********************************************************************************
```

Figure A.4: Solution re1 CLMethod=Direct, CRfun=H1, SimFun=CorrCoef.



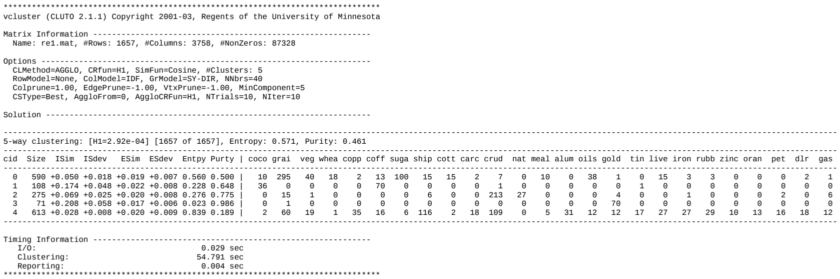
```
********************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----------------------------------------------------------
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options ----------------------------------------------------------------------
  CLMethod=Direct, CRfun=H1, SimFun=CorrCoef, #Clusters: 5
  RowModel=None, ColModel=None, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H1, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
5-way clustering: [H1=4.66e-04] [1657 of 1657], Entropy: 0.633, Purity: 0.414
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | coco grai  veg whea copp coff suga ship cott carc crud  nat meal alum oils gold  tin live iron rubb zinc oran  pet  dlr  gas
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
  0   213 +0.224 +0.067 +0.061 +0.021 0.752 0.338 |    3   72    6    2    4    5   11   12    5    4   37    5    5    1    5    5    1    7   10    0    2    1    5    0    5
  1   325 +0.182 +0.054 +0.047 +0.016 0.624 0.415 |    7  135   18   15    2    8   71    6    1    4    0    1    9    9   21    2    5    4    1    3    1    0    1    0    1
  2   313 +0.169 +0.048 +0.044 +0.017 0.262 0.796 |    0    2   31    0    0    1    0    5    0    0  249    8    0    0    7    2    0    0    0    0    0    0    1    0    7
  3   400 +0.080 +0.031 +0.039 +0.018 0.743 0.300 |   38  120    4    1   14   79   16    4    7    4    2    3    1    7    6    3    5   25    9   24    3    9    7    8    1
  4   406 +0.046 +0.019 +0.027 +0.015 0.756 0.271 |    0   42    1    1   17    6    8  110    4    8   42   10    0   14   11   75    7    6   11    5    4    3    4   12    5
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information -----------------------------------------------------------
   I/O:                                   0.026 sec
   Clustering:                            9.332 sec
   Reporting:                             0.129 sec
********************************************************************************
```

Figure A.5: Solution re1 CLMethod=AGGLO, CRfun=I2, SimFun=Cosine.



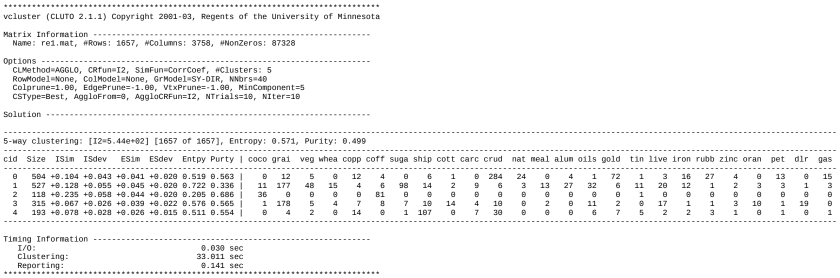
```
********************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----------------------------------------------------------
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options ----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=I2, SimFun=Cosine, #Clusters: 5
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
5-way clustering: [I2=3.84e+02] [1657 of 1657], Entropy: 0.564, Purity: 0.482
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | coco grai  veg whea copp coff suga ship cott carc crud  nat meal alum oils gold  tin live iron rubb zinc oran  pet  dlr  gas
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
  0   268 +0.046 +0.018 +0.017 +0.007 0.538 0.373 |    0   27    2    0    3    0    2  100    0    7   47    0    2    0    6   64    0    2    0    0    1    0    3    0    2
  1   237 +0.071 +0.034 +0.020 +0.008 0.631 0.354 |   40   11    4    0   27   84    5    4    2    1    5    1    0    0    1    7   17    0    0   27    1    0    0    0    0
  2   364 +0.058 +0.021 +0.020 +0.008 0.394 0.701 |    0   17   16    0    1    0    0    3    0    1  255   26    0    8    5    6    1    1    2    0    0    0    7    0   15
  3   392 +0.055 +0.020 +0.021 +0.008 0.457 0.673 |    5  264    9   19    4    6    3   12   14    3    4    0    8    0   21    0    0   11    2    2    1    0    0    2    2
  4   396 +0.045 +0.020 +0.022 +0.009 0.805 0.242 |    3   52   29    0    2    9   96   18    1    8   19    0    5   23   17   10    0   28   27    3    7   13    8   18    0
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information -----------------------------------------------------------
   I/O:                                   0.033 sec
   Clustering:                            0.575 sec
   Reporting:                             0.004 sec
********************************************************************************
```

Figure A.6: Solution re1 CLMethod=AGGLO, CRfun=H1, SimFun=Cosine.



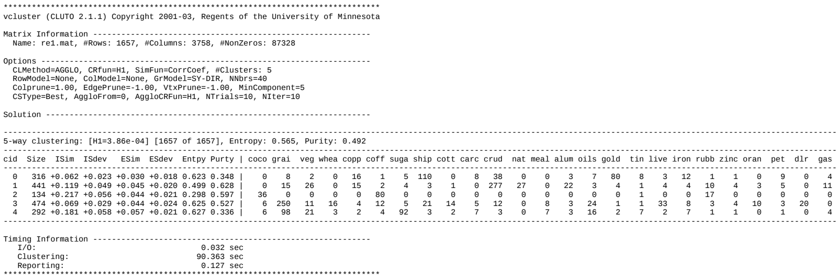
```
********************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----------------------------------------------------------
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options ----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=H1, SimFun=Cosine, #Clusters: 5
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H1, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
5-way clustering: [H1=2.92e-04] [1657 of 1657], Entropy: 0.571, Purity: 0.461
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | coco grai  veg whea copp coff suga ship cott carc crud  nat meal alum oils gold  tin live iron rubb zinc oran  pet  dlr  gas
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
  0   590 +0.050 +0.018 +0.019 +0.007 0.560 0.500 |   10  295   40   18    2   13  100   15   15    2    7    0   10    0   38    1    0   15    3    3    0    0    0    2    1
  1   108 +0.174 +0.048 +0.022 +0.008 0.228 0.648 |   36    0    0    0    0   70    0    0    0    0    1    0    0    0    0    0    1    0    0    0    0    0    0    0    0
  2   275 +0.069 +0.025 +0.020 +0.008 0.276 0.775 |    0   15    1    0    0    0    0    6    0    0  213   27    0    0    0    4    0    0    1    0    0    0    2    0    6
  3    71 +0.208 +0.058 +0.017 +0.006 0.023 0.986 |    0    1    0    0    0    0    0    0    0    0    0    0    0    0    0   70    0    0    0    0    0    0    0    0    0
  4   613 +0.028 +0.008 +0.020 +0.009 0.839 0.189 |    2   60   19    1   35   16    6  116    2   18  109    0    5   31   12   12   17   27   27   29   10   13   16   18   12
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information -----------------------------------------------------------
   I/O:                                   0.029 sec
   Clustering:                           54.791 sec
   Reporting:                             0.004 sec
********************************************************************************
```

Figure A.7: Solution re1 CLMethod=AGGLO, CRfun=I2, SimFun=CorrCoef.



```
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----------------------------------------------------------
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options ----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=I2, SimFun=CorrCoef, #Clusters: 5
  RowModel=None, ColModel=None, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
5-way clustering: [I2=5.44e+02] [1657 of 1657], Entropy: 0.571, Purity: 0.499
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | coco grai  veg whea copp coff suga ship cott carc crud  nat meal alum oils gold  tin live iron rubb zinc oran  pet  dlr  gas
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
  0   504 +0.104 +0.043 +0.041 +0.020 0.519 0.563 |    0   12    5    0   12    4    0    6    1    0  284   24    0    4    1   72    1    3   16   27    4    0   13    0   15
  1   527 +0.128 +0.055 +0.045 +0.020 0.722 0.336 |   11  177   48   15    4    6   98   14    2    9    6    3   13   27   32    6   11   20   12    1    2    3    3    1    3
  2   118 +0.235 +0.058 +0.044 +0.020 0.205 0.686 |   36    0    0    0    0   81    0    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0
  3   315 +0.067 +0.026 +0.039 +0.022 0.576 0.565 |    1  178    5    4    7    8    7   10   14    4   10    0    2    0   11    2    0   17    1    1    3   10    1   19    0
  4   193 +0.078 +0.028 +0.026 +0.015 0.511 0.554 |    0    4    2    0   14    0    1  107    0    7   30    0    0    0    6    7    5    2    2    3    1    0    1    0    1
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information -----------------------------------------------------------
   I/O:                                     0.030 sec
   Clustering:                             33.011 sec
   Reporting:                               0.141 sec
*******************************************************************************
```

Figure A.8: Solution re1 CLMethod=AGGLO, CRfun=H1, SimFun=CorrCoef.



```
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----------------------------------------------------------
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options ----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=H1, SimFun=CorrCoef, #Clusters: 5
  RowModel=None, ColModel=None, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H1, NTrials=10, NIter=10

Solution ---------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
5-way clustering: [H1=3.86e-04] [1657 of 1657], Entropy: 0.565, Purity: 0.492
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | coco grai  veg whea copp coff suga ship cott carc crud  nat meal alum oils gold  tin live iron rubb zinc oran  pet  dlr  gas
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
  0   316 +0.062 +0.023 +0.030 +0.018 0.623 0.348 |    0    8    2    0   16    1    5  110    0    8   38    0    0    3    7   80    8    3   12    1    1    0    9    0    4
  1   441 +0.119 +0.049 +0.045 +0.020 0.499 0.628 |    0   15   26    0   15    2    4    3    1    0  277   27    0   22    3    4    1    4    4   10    4    3    5    0   11
  2   134 +0.217 +0.056 +0.044 +0.021 0.298 0.597 |   36    0    0    0    0   80    0    0    0    0    0    0    0    0    0    0    1    0    0   17    0    0    0    0    0
  3   474 +0.069 +0.029 +0.044 +0.024 0.625 0.527 |    6  250   11   16    4   12    5   21   14    5   12    0    8    3   24    1    1   33    8    3    4   10    3   20    0
  4   292 +0.181 +0.058 +0.057 +0.021 0.627 0.336 |    6   98   21    3    2    4   92    3    2    7    3    0    7    3   16    2    7    2    7    1    1    0    1    0    4
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information -----------------------------------------------------------
   I/O:                                     0.032 sec
   Clustering:                             90.363 sec
   Reporting:                               0.127 sec
*******************************************************************************
```