

1 General Picture

In this part of the project, you will conduct some experiments using different clustering algorithms and different datasets. Major project goals are,

- Use the popular K-means and DBSCAN clustering algorithms
- Analyze the results of these algorithms using the Cougar² framework and some of its region discovery functions; namely, you will learn what datasets, fitness functions, interestingness measures in the framework are - in Part3 you will implement a clustering algorithm yourself.
- Run a region discovery algorithm called CLEVER.
- Visualize and analyze results for running the above three algorithms for 2 datasets.

2 What Will Be Used in the Project?

- Clustering Algorithms: DBSCAN and K-Means (You will use WEKA for these algorithms. Weka can be downloaded from Weka 3.5.6 <http://www.cs.waikato.ac.nz/ml/weka/>)
- Datasets: 9- Diamonds, Arsenic, Complex-9 (optional)
- Fitness Functions: Purity for the first dataset and Variance for the second one
- You will also use Mean Squared Error (MSE) for both the datasets.
- Visualization: Microsoft Excel or Matlab

Datasets and a document on 'How to run an Experiment?' can be found at: <http://www2.cs.uh.edu/~ykuo/>

3 Project Tasks

1. Run the K-means clustering algorithm with $K=15$, $K=9$ (twice), $K=6$ for the datasets tested. Cluster only using the spatial attributes and ignore non-spatial attributes!
2. Run DBSCAN several times for the datasets to determine the best parameter setting for MinPoints and ϵ . Report the best parameter settings for each dataset. Cluster based on the spatial attributes and ignoring other attributes! Save the best two results you obtained for the two datasets.
3. Visualize the results obtained in steps 1 and 2 (the 4 k-means results, and the two best results obtained by DBSCAN). Interpret the results!

4. Compute the fitness of the obtained clustering using the MSE, purity, and variance fitness measures. Report the 3 regions that have the highest interestingness; report the 3 regions that have lowest fitness. Use purity for 9-Diamonds and variance for Arsenic Dataset. Use MSE for both the datasets. While using MSE use the actual MSE value and not the fitness value returned by the interestingness measure.
5. Run the CLEVER algorithms with the purity fitness function (Parameters for CLEVER: $\beta = 1.01, \hat{k} = 50, p = 50, q = 50$ *NeighborhoodSize* = 3, $p(\text{insert}) = 0.2, p(\text{delete}) = 0.2, p(\text{replace}) = 0.6$ and Parameters for Purity: $\eta = 2, th = 0.6$) for the 9-Diamond dataset, and visualize and compare the results with those obtained by the other two algorithms.
 - (a) You can also run CLEVER with Purity fitness function on Complex-9 dataset.
6. Run the CLEVER algorithms with the variance fitness function (Parameters for CLEVER: $\beta = 1.2, \hat{k} = 40, p = 50, q = 50, \text{NeighborhoodSize} = 3, p(\text{insert}) = 0.2, p(\text{delete}) = 0.2, p(\text{replace}) = 0.6$ and Parameters for Variance: $\eta = 1, b = 10$) for the Arsenic dataset, and visualize and compare the results with those obtained by the other two algorithms.
7. Submit detailed reports that summarize your findings.

4 Region Discovery Framework

A spatial dataset consists of spatial and non-spatial (meta) attributes. A fitness function that evaluates a clustering depends only on meta attributes and the discovered clustering. The fitness function $q(X)$ depends on a clustering $X = \{r_1, r_2, \dots, r_k\}$. Each region $r_i \subseteq O$ is a subset of the dataset O . Regions are mutually exclusive, $r_i \cap r_j = \phi$ for $i \neq j$, and mutually exhaustive, $r_1 \cup r_2 \cup \dots \cup r_k \subseteq O$. The clustering and some additional summary information is the output of a region discovery algorithm. We consider a subclass of additive fitness functions:

$$q(X) = \sum_{r \in X} q(r) = \sum_{r \in X} i(r) |r|^\beta$$

where $q(r)$ is the region specific reward function, $|r|$ is the size of the region, and the exponent $\beta > 1$ weights the size of the cluster. The function $i(r)$ is an interestingness measure defined on a region $r \in X$.

5 Design

Figure 1: Class diagram for fitness functions and interestingness measures in Region Discovery Framework

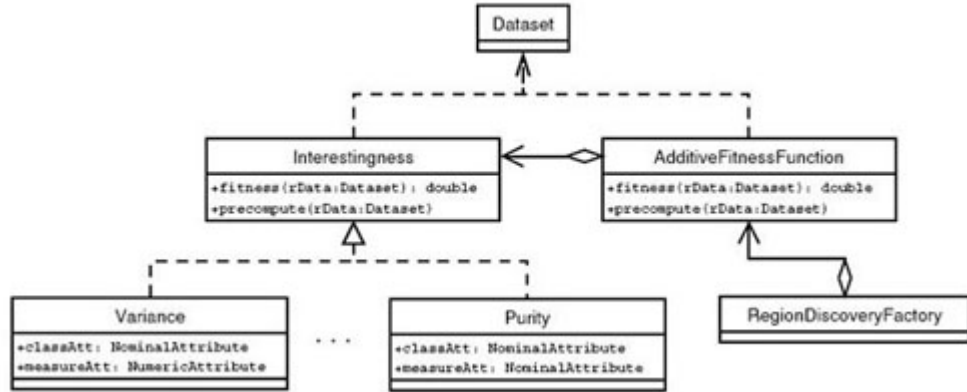
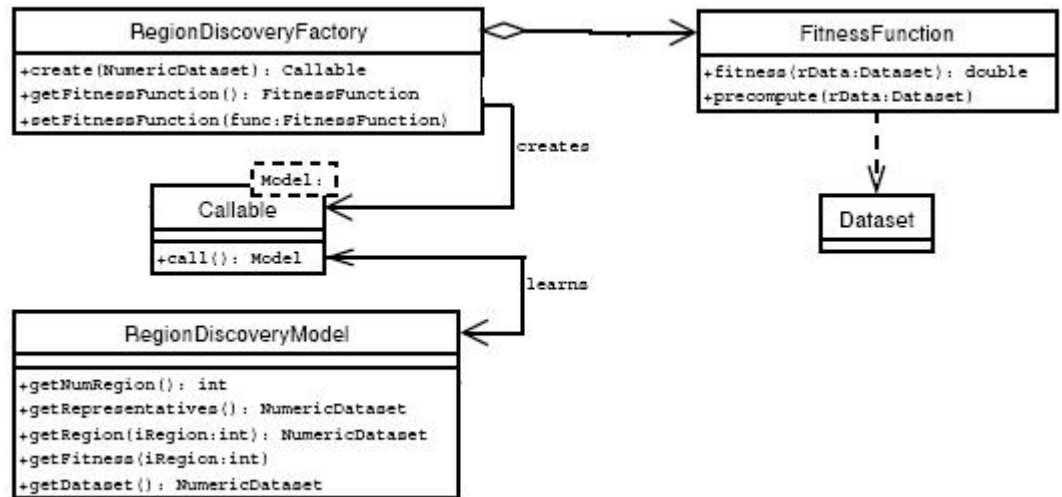


Figure 2: Class diagram for Region Discovery Algorithm classes



6 Measures of Interestingness

For simplicity we assume that the argument is a Dataset object and this Dataset corresponds to a single region.

6.1 Mean Squared Error (MSE)

Let μ be the mean of the values in the dataset. The measure of interestingness is defined as:

$$\begin{aligned}
i(r) &= \begin{cases} 0 & \text{if } \left(\frac{1}{MSE+1}\right) < th \\ \left(\left(\frac{1}{MSE+1}\right) - th\right)^\eta & \text{otherwise} \end{cases} \\
MSE(r) &= \frac{1}{m} \sum_{o \in r} \|r - centroid(r)\|^2 \\
centroid(r) &= \frac{1}{m} \sum_{o \in r} r
\end{aligned}$$

In this measure, only the real data is used and not the meta-data. All data attributes are numeric attributes. $\eta > 1$ is the scaling factor. $th > 0$ is the threshold. m is the size of the dataset. Euclidian distance is used for the distance calculation.

6.2 Purity

Purity is defined as fraction of examples that belong to the majority class. The majority class is the class with the most number of examples. In the event that two classes have an equal number of examples, the majority class is arbitrarily selected.

$$i(r) = \begin{cases} 0 & \max_c p_c(r) < th \\ (\max_c p_c(r) - th)^\eta & \text{otherwise} \end{cases}, c \in cl(O)$$

Attribute used is the nominal class attribute. $\eta > 1$ is the scaling factor. $cl(O)$ is the set of classes in the dataset. $th > 0$ is the threshold.

6.3 Variance

The sample variance of a numeric meta-attribute m :

$$\begin{aligned}
i(r) &= \left(\left(\begin{cases} 0 & Var(r(m)) < Var(O(m)) \\ \min\left\{1, \log_b \frac{Var(r(m))}{Var(O(m))}\right\} & \text{otherwise} \end{cases} \right) \right)^\eta \\
Var(r(m)) &= \frac{1}{n-1} \sum_{o \in r} (x_m - \mu_m)^2
\end{aligned}$$

Where n is the number of examples, x_m is the value of the m th meta-attribute in x , and μ is the mean value of the m th meta-attribute. $b > 1$ is the base of the logarithm. $\eta > 0$ is the scaling factor. $O(m)$ is the m th meta-attribute of the dataset and $Var(O(m))$ is the variance of the m th meta-attribute in the dataset.

7 Coding Conventions

The code that you will write for the experiments should follow a few programming conventions:

- `_X` X is a private member variable.
- `iX` is an index to the array or collection named X.
- `tY` Y is the value you are testing as in: `tSize = _testObject.size();`
- `eX` X is the known value against which you are comparing as in: `eSize = 7; assertEquals(eSize, tSize);`
- `fX` is an element of the array named X whose index is `iX` as in: `fX = X[iX];`
- `uX` X is the return value of an method as in: `uSum = 1+2; return uSum;`
- `nX` length of the array X as in: `nX = X.length;` or `nX = X.size();` or `nX = X.getNumMetaAttributes();`
- `rX` Read-only reference to X in method signature
- `wX` Read-Write reference to X in method signature
- `whatXDoes` The name of a variable should indicate how it is used in the function. Avoid single letter names as in (bad): `z = x[y]` could better be written `fCluster = cluster[iCluster]`

8 Submission

Email the following to ykuo@cs.uh.edu (notice different deadlines for the different tasks of the project):

1. Submit detailed report for the tasks 1-4 by October 6, 11:00 PM CST.
2. Submit detailed report for the tasks 5 and 6 by October 11, 11:00 PM CST.