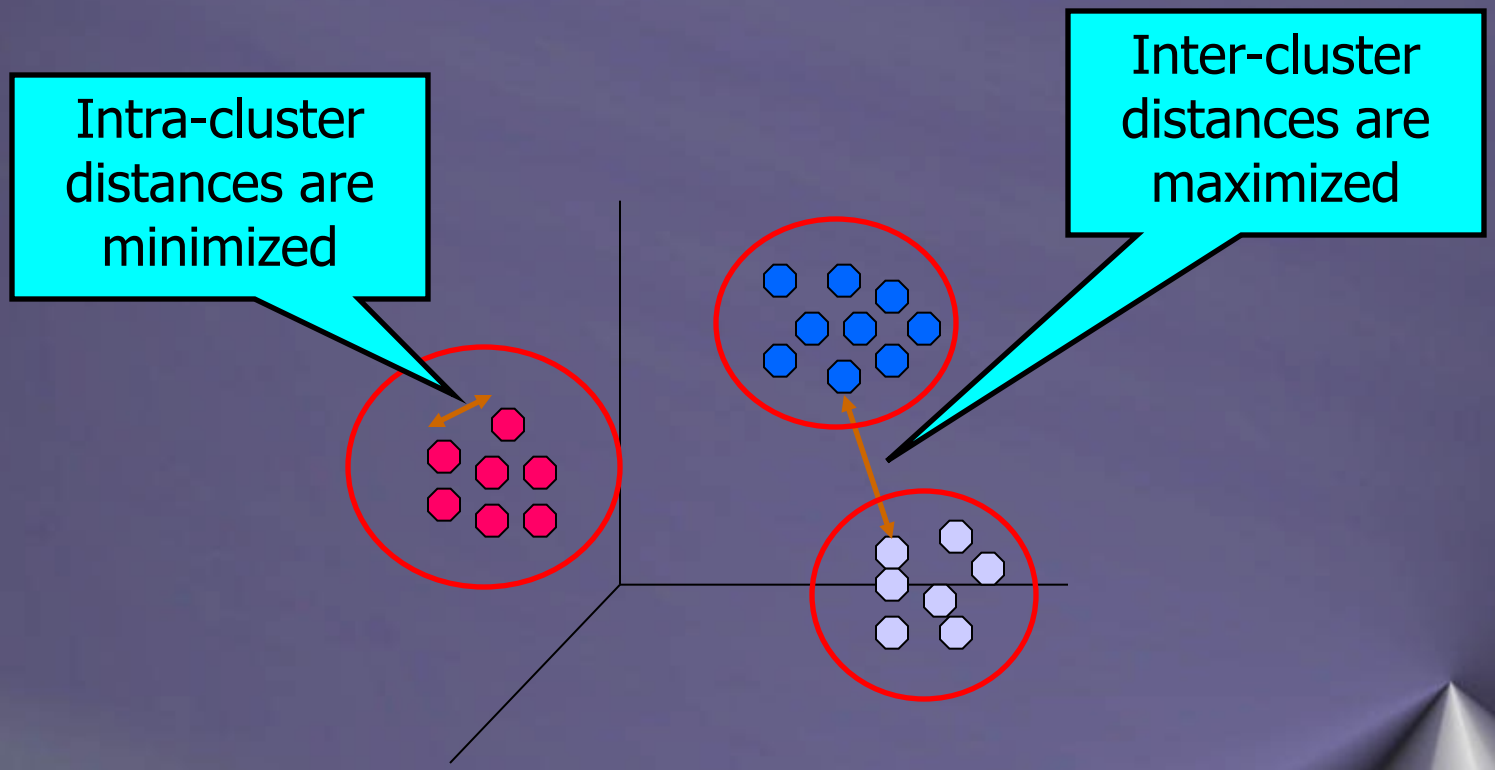
Clustering

Cristóbal Romero
Department of Computer Science and Numerical Analysis
University of Cordoba, Spain
cromero@uco.es

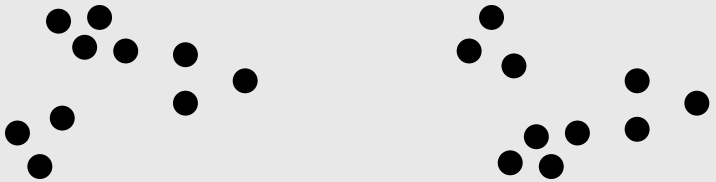
KDIS Lab
KNOWLEDGE DISCOVERY AND INTELLIGENT SYSTEMS

What is Cluster Analysis?

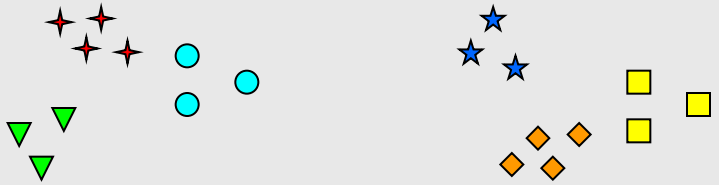
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



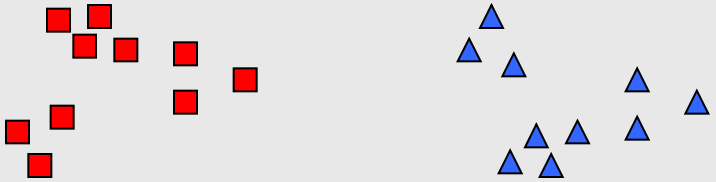
Notion of a Cluster can be Ambiguous



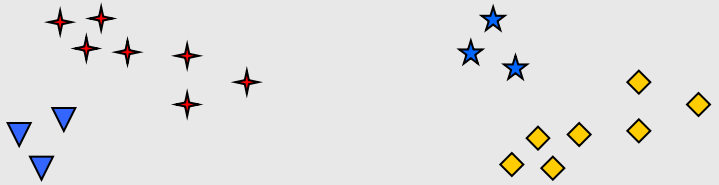
How many clusters?



Six Clusters



Two Clusters



Four Clusters

Clustering Algorithms

- Considerable progress has been made in clustering methods:
  - Partitioning: k-means, k-medoids, CLARANS
  - Hierarchical: BIRCH, ROCK, CHAMELEON
  - Density-based: DBSCAN, OPTICS, DenClue
  - Grid-based: STING, WaveCluster, CLIQUE
  - Model-based: EM, Cobweb, SOM
  - Frequent pattern-based: pCluster
  - Constraint-based: COD, constrained-clustering

Typical Alternatives to Calculate the Distance between Clusters

- Single link: smallest distance between an element in one cluster and an element in the other, i.e., $dis(K_i, K_j) = min(t_{ip}, t_{jq})$
- Complete link: largest distance between an element in one cluster and an element in the other, i.e., $dis(K_i, K_j) = max(t_{ip}, t_{jq})$
- Average: avg distance between an element in one cluster and an element in the other, i.e., $dis(K_i, K_j) = avg(t_{ip}, t_{jq})$
- Centroid: distance between the centroids of two clusters, i.e., $dis(K_i, K_j) = dis(C_i, C_j)$
- Medoid: distance between the medoids of two clusters, i.e., $dis(K_i, K_j) = dis(M_i, M_j)$
  - Medoid: one chosen, centrally located object in the cluster

Clustering for predicting student performance from forum data

- The more students participate in the forum for a certain course, the more involved they will be in the subject matter of that course.
- Following this line, in this study we try to test whether or not there is a correlation between the participation of students in Moodle forums and their final course marks.

Background

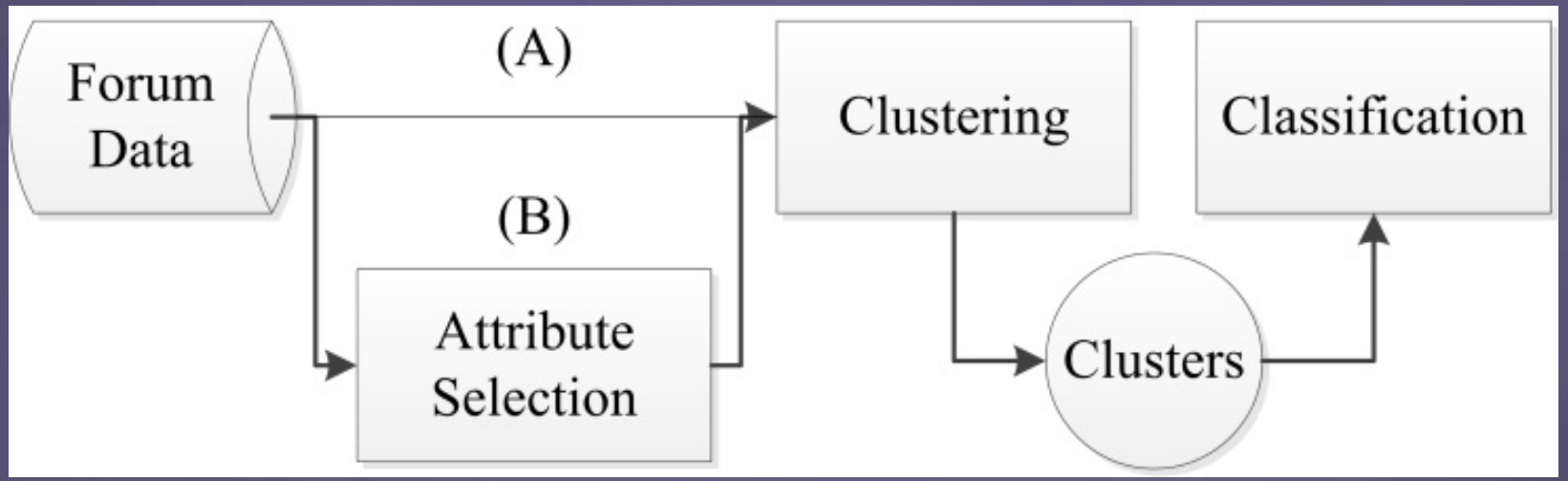
- The use of data mining is a potential strategy for discovering and building alternative representations for the data underlying discussion forums.
- There is less published work on the use of data mining to predict student performance based on forum usage data.
- Furthermore, the use of clustering for classification has not yet been applied in an educational context.

Proposed Approach

- We propose to use a meta-classifier that uses a cluster for classification approach based on the assumption that each cluster corresponds to a class.
- For all cluster algorithms, the number of clusters generated is the same as the number of class labels in the dataset. We use this approach to test if student participation in forums is related to whether they pass or fail the course.

Proposed Approach

- Proposed classification via clustering approach



Description of the data used

- The dataset used in this work was gathered from a Moodle forum used by university students during a first-year course in computer engineering in 2011.
- We developed a new module for Moodle specifically to obtain a summary dataset file.

Student	nMessages	nThreads	nReplies	nWords	nSentences	nR
[illegible]	3	0	3	67	3	
[illegible]	6	1	5	513	1	
[illegible]	1	1	0	17	2	
[illegible]	2	0	2	43	2	

Description of the data used

- Some forum statistics are:

Number of students	Number of messages	Number of threads	Number of replies
114	1014	81	933

- The variables relating to forum usage are:

Attribute	Description
nMessages	Number of messages sent per student
nThreads	Number of threads created per student
nReplies	Number of replies sent per student
nWords	Number of words written by student
nSentences	Number of sentences written by student
nReads	Number of messages read on the forum
tTime	Total time, in hours, spent on forum
aEvaluation	Average score of the messages
dCentrality	Degree centrality of the student
dPrestige	Degree prestige of the student
fMark	Final mark obtained by the student

Experimental Results

- In the first experiment, we executed the following clustering algorithms provided by Weka for classification via clustering using all attributes: EM, FarthestFirst, Xmeans, sIB HierarchicalClusterer and SimpleKMeans.
- In the second experiment, we repeated all the previous executions selecting the best attributes, based on the assumption that not all the available attributes are discriminative factors in the final marks.

Experimental Results

- We apply a range of feature-selection algorithms. To rank the attributes, we counted the number of times each attribute was selected by each attribute-selection algorithm.
- We selected as the best attributes the first six attributes in the ranking, because these were selected by at least half of the algorithms.

Experimental Results

Attribute	Frequency
dCentrality	9
nMessages	8
nReplies, nWords	7
dPrestige	6
aEvaluation	5
nSentences, nReads, nThreads	3
tTime	1

Experimental Results

- The table shows the overall accuracy (rate of correctly classified students) using all the available attributes (A) and using only the six selected attributes (B).

Clustering algorithm	**(A)**	**(B)**
EM	**0.842**	**0.894**
FarthestFirst	0.526	0.535
HierarchicalClusterer	0.578	0.570
sIB	0.710	0.578
SimpleKMeans	0.666	0.640
Xmeans	0.666	0.640

Experimental Results

- In the third experiment, we compared the accuracy of the previous classification via clustering approach with that of traditional classification algorithms by executing a representative number of classifications of different types: Rules-based algorithms, Trees-based algorithms, Functions-based algorithms and Bayes-based algorithms.

Experimental Results

Algorithms	(A)	(B)
DTNB	0.859	0.833
JRip	0.833	0.815
NNge	0.842	0.807
Ridor	0.833	0.842
ADTree	0.859	0.842
J48	0.824	0.807
LADTree	0.868	0.850
RandomForest	0.850	0.833
Logistic	0.859	0.850
MultilayerPerceptron	0.842	0.868
RBFNetwork	0.868	0.886
SMO	0.868	0.886
BayesNet	**0.877**	0.842
NaiveBayesSimple	0.859	**0.894**

Experimental Results

- Finally, we show the cluster centroids for the EM algorithm when using the six selected attributes that have yielded the best accuracy.

Attributes	Cluster 0	Cluster 1
nMessages	1.2199	14.8905
nReplies	1.1599	13.6718
nWords	18.4599	668.8039
aEvaluation	0	0.7751
dCentrality	0.0011	0.1565
dPrestige	0	0.1021

Conclusions

- Based on the results obtained using several clustering and classification algorithms, we can answer the two initial questions.
- a) Yes, student participation in the course forum was a good predictor of the final marks for the course.
- b) Yes, the proposed classification via clustering approach obtained similar accuracy to traditional classification algorithms using our forum data.

The end

Thank you for your interest.

Questions? Comments?