Structural Clustering and Visualization for Multi-Objective Decision Making

Yong-Seog Kim
Utah State University

Follow this and additional works at: http://aisel.aisnet.org/amcis2004

Recommended Citation
http://aisel.aisnet.org/amcis2004/240

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2004 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
Structural Clustering and Visualization for Multi-Objective Decision Making

YongSeog Kim
Utah State University
Business Information Systems
Logan, UT 84341
ykim@b202.usu.edu

ABSTRACT
Clustering is an important problem in knowledge discovery and decision making. In this study, a multi-objective genetic algorithm (MOGA) is used to search for well separated clusters and each evolved solution is evaluated by both data-driven and human-driven metrics developed for this study. The proposed system in this paper also allows the decision maker to navigate non-dominated solutions and to choose one of them as the final solution. Experimental results on both synthetic and real data sets show promise in finding non-dominated solutions while exploring more promising objective space given the same amount of computational time.

Keywords
Clustering, genetic algorithm, visualization, pareto optimization

INTRODUCTION
Clustering is the process of partitioning records into a fixed number of groups (or clusters) based on similarity or dissimilarity among records and has been recently used for profiling Web usage (Anandarajan, 2002) and data analysis in a data warehousing environment (Ester et al., 1998). However, many researchers in previous research use the complete set of input variables or a pre-selected subset of features in their clustering analyses without considering interactions between input variables (or features) and clustering results. Further, most clustering results have been evaluated with only data-driven metrics such as intra-cluster compactness (how similar the elements of each cluster are) and inter-cluster separability (how dissimilar the clusters are). Therefore, it is possible that clustering results with the highest value of data-driven metrics may lack of practical and managerial implications.

The ultimate goal of information systems for clustering and other data analyses is to provide decision makers with information that is easy to understand and useful for making a right decision at the right time. It is also highly recommended that decision support systems utilize all available expertise and experiences of human experts to find solutions for real world applications. Therefore, there is a strong need to evaluate clustering results using not only data-driven but also human-driven metrics. By evaluating outputs with human-driven metrics, the proposed system can provide solutions that reflect decision makers’ subjective preference on managerial instruments. Further, decision support systems should be able to incorporate multiple and often conflicted objectives that decision makers might have. The proposed system considers a clustering problem from the perspective of a multi-objective or Pareto optimization. In a Pareto optimization, a solution is said to dominate another if it has higher values along all the objective functions. The Pareto front is defined as the set of non-dominated solutions and the ultimate goal of the proposed system is to approximate as best possible the Pareto front, presenting the decision maker with a set of high-quality compromise solutions from which to choose.

The interaction between the decision maker and the model is a critical success factor of the proposed system. The interaction between two parties occurs when the decision maker provides the model a relevance map that reflects the decision maker’s subjective weights on the relevance or importance of input variables in decision making process. Since the relevance map from the decision maker affects the scope of search space and the structure of resulting clustering models, clustering with the relevance map is termed as structural clustering in this study. The decision maker also interacts with the system to make the final decision. By providing a graphical decision-aid tool, the proposed model intends to help the decision maker easily make alternative decisions based on currently available resources and options.
K-MEANS ALGORITHM AND HEURISTIC METRICS

K-means (Duda and Hart, 1973) is one of the most often used nonhierarchical clustering methods because of its simplicity. Starting with a random initial partition, K-means iteratively assigns each data point to the cluster whose centroid is located nearest to the given point, and recalculates the centroids based on the new set of assignments until no points are reassigned. In this study, four fitness criteria are used to evaluate cluster quality after each objective is normalized into the unit interval. Two data-driven metrics—\( F_{\text{within}} \) and \( F_{\text{complexity}} \)—and two human-driven metrics—\( F_{\text{relevance}} \) and \( F_{\text{interest}} \)—are described as follows:

- \( F_{\text{within}} \): This metric is meant to favor dense clusters by measuring cluster cohesiveness. Formally, let \( x_i, i = 1, \ldots, n \), be data points and \( x_{ij} \) be the value of the \( j \)-th feature of \( x_i \). Let \( d \) be the dimension of the selected feature set, \( J \), and \( K \) be the number of clusters. The cluster membership variables \( \alpha_{ik} \) is defined as follows:
  \[
  \alpha_{ik} = \begin{cases} 
  1 & \text{if } x_i \text{ belongs to cluster } k \\
  0 & \text{otherwise}
  \end{cases}
  \]
  where \( k = 1, \ldots, K \) and \( i = 1, \ldots, n \). The centroid of the \( k \)-th cluster, \( \gamma_k \), is a set of its coordinates defined for each dimension \( j \in J \) as follows:
  \[
  \gamma_j^k = \frac{\sum_{i=1}^{n} \alpha_{ik} x_{ij}}{\sum_{i=1}^{n} \alpha_{ik}}, \quad j \in J.
  \]
  \( F_{\text{within}} \) can finally be computed as follows:
  \[
  F_{\text{within}} = \frac{1}{Z_w} \frac{1}{d} \sum_{k=1}^{K} \sum_{i=1}^{n} \alpha_{ik} \sum_{j \in J} (x_{ij} - \gamma_j^k)^2
  \]
  where the normalization by the number of selected features \( d \) compensates for the dependency of the distance metric on the feature subspace dimensionality. \( Z_w \) is a normalization constant meant to achieve \( F_{\text{within}} \) values spanning the unit interval and its value is set empirically by running preliminary experiments.

- \( F_{\text{complexity}} \): This metric is aimed at minimizing the number of selected features for clustering and defined as follows:
  \[
  F_{\text{complexity}} = 1 - \frac{d - 1}{D - 1}
  \]
  where \( D \) is the dimensionality of the full data set. It is expected that lower complexity (i.e., few chosen features) will lead to easier interpretability of solutions as well as better generalization.

- \( F_{\text{relevance}} \): This metric is meant to favor clusterings with a subset of features that decision makers believe important and, therefore, want to have in their final decision making. The presented system in this paper takes a relevance map that the decision maker shows her subjective preferences over each feature using boolean or real values. Clustering results with more relevant features will be preferred to clustering results with less relevant features. Formally, \( F_{\text{relevance}} \) is defined as follows:
  \[
  F_{\text{relevance}} = \frac{\sum_{i=1}^{d} \beta_i}{d}
  \]
  where \( \beta_i \) represents the relevance value of feature \( i \) specified by the decision maker and its value is between 0 and 1.

- \( F_{\text{interest}} \): Although decision makers guide search direction toward more promising area by specifying their expertise and preferences through \( F_{\text{relevance}} \), it is possible that there exist hidden and interesting clusters that human experts do not expect in advance. This becomes very important from the perspective of knowledge discovery. In order to look for “surprising” clusters, a new metric, \( F_{\text{interest}} \), is defined as follows:
  \[
  F_{\text{interest}} = \frac{F_{\text{within}}}{F_{\text{relevance}} + 1.0}
  \]
  Intuitively, dense clusters (i.e., high value of \( F_{\text{within}} \)) constructed mostly by irrelevant features (low value of \( F_{\text{relevance}} \)) will be very interesting because human experts do not expect to see very well-separated and cohesive clusters using chosen irrelevant features.
MULTI-OBJECTIVE GENETIC ALGORITHMS (MOGA)

Decision Making with MOGA versus Standard GA

Since (Goldberg, 1989), various types of GAs have been used for many different applications. Typically, a GA starts with and maintains a population of chromosomes (or agents) that correspond to solutions to the problem. Since a GA cannot typically search the whole space of the solution space, it gradually limits its focus to more highly fit regions of the search space. Although standard GAs have been widely applied for various applications, they have shown limited capability when decision makers should take into account multiple, conflicted objectives simultaneously. In order to provide a clear picture of the tradeoffs among the various objectives, a number of multi-objective extensions of GAs have been proposed in recent years (Horn, 1997) and the problem has been formulated as a multi-objective or Pareto optimization.

In a Pareto optimization, the goal is to approximate as best possible the Pareto front defined as the set of nondominated solutions, presenting the decision maker with a set of high-quality solutions from which to choose. By providing a set of alternative solutions to the decision maker, the proposed system helps her to choose the right solution for a specific application. The decision maker can select a final model after determining her relative preferences of criteria for application. As a multi-objective extension of GAs, this study uses evolutionary local selection algorithm (ELSA) that has been successfully applied for a customer targeting problem (Kim et al., 2003). For a more extensive discussion of the algorithm, please refer to (Menczer et al., 2000).

Structure of ELSA/K-means Model

The components of the research model and the information flow among components are shown in Figure 1. One of the most important features of the proposed system is the fact that the decision maker actively interacts with the system. The system first takes a relevance map from the decision maker. The relevance map is a set of relevance values that the decision maker subjectively determines based on her expertise or preferences. The relevance value of each feature can be specified using either boolean values (1 for relevant features, 0 otherwise) or real values between 0 and 1 (higher value for highly relevant features). Once the relevance map is presented to the system, the search and evaluation components take their places. The main purpose of ELSA is to explore combinatorial solution space of a set of feature subsets and the appropriate number of clusters. Each agent in ELSA corresponds to a candidate solution with a specific bit string to represent feature subset and the number of clusters to build. Once a candidate solution is passed to a K-means clustering algorithm, the corresponding bit string is parsed into a feature subset $J$ and a cluster number $K$. Given the projection of the data set onto $J$, K-means forms $K$ clusters and returns the four fitness criteria—$F_{\text{within}}, F_{\text{complexity}}, F_{\text{relevances}},$ and $F_{\text{interest}}$—to ELSA.
Based on the returned metric values, ELSA biases its search direction toward more promising regions. This routine continues until the maximum number of iterations is attained. Once the maximum number of iterations is attained, all evaluated solutions are passed to the visualization component for further analysis. The visualization component finds nondominated solutions among all evaluated solutions and visually presents them to the decision maker. The visualization tool allows the decision maker to navigate non-dominated solutions so that she can choose one of them as the final solution after subjectively weighting multiple objectives.

**EXPERIMENTAL RESULTS**

**Experiment 1**

In order to evaluate the proposed model, a moderate-dimensional synthetic data set is constructed. The data set has $N = 500$ points and $D = 30$ features. The feature set consists of “significant” features, “Gaussian noise” features, and “white noise” features. It is constructed so that the first 10 features are significant, with 5 “true” normal clusters consistent across these features. The next 10 features are Gaussian noise, with points randomly and independently assigned to 2 normal clusters along each of these dimensions. The remaining 10 features are white noise in which points are drawn from uniform distributions. The standard deviation of the normal distributions is $\sigma \approx 0.06$ and the means are themselves drawn from uniform distributions in the unit interval, so that the clusters may overlap. The Figure 2 presents some 2-dimensional projections of the synthetic data set. The following three models are compared on the synthetic data set.

- **Model 1 (Clustering without a relevance map):** This model uses only two data-driven metrics ($F_{within}$ and $F_{complexity}$) to evaluate the quality of clusters.

- **Model 2 (Clustering with a boolean value relevance map):** In this model, the decision maker assigns a boolean value—0 for not relevant and 1 for relevant feature—to each input variable based her subjective confidence in relevance of each variable. In this study, the first ten features have the relevance value of one and the other 20 features have the relevance value of zero.

- **Model 3 (Clustering with a real value relevance map):** This model is the same as Model 2 except that it allows the decision maker to assign any real value between 0 and 1 as the relevance value of each feature. To implement real value relevance scheme on the synthetic data, relevance values of 0.75, 0.5, and 0.25 are assigned to the first ten, the middle ten, and the last ten features, respectively.

![Figure 2. A few 2-dimensional projections of the synthetic data set](image)

It is expected that ELSA with boolean and real value relevance maps will be able to explore more promising solutions with high $F_{within}$ values than ELSA without relevance maps. This is mainly because ELSA receives a good feedback from high $F_{relevance}$ values whether or not candidate solutions are relevant. To graphically show the performance of three different models, candidate fronts of three models are constructed and shown in Figure 3. The term candidate front is used for the set...
of solutions with the highest $F_{\text{within}}$ value at every $F_{\text{complexity}}$ value among all candidate solutions. It is clear from Figure 3 that ELSA with either one of the relevance maps explores solutions with higher $F_{\text{within}}$ values where $F_{\text{complexity}} > 0.66$. The solution space where $F_{\text{complexity}} > 0.66$ is very interesting because there are 10 significant features in the synthetic data and when 10 features are used for clustering, its $F_{\text{complexity}}$ becomes $1 - (10-1)/(30-1) \approx 0.6896$. In order to numerically measure the difference in the performance of three models, the coverage of model $k$ from a candidate front is computed as follows:

$$\text{coverage}^k = \sum_{i \in F_{\text{complexity}}} F_{\text{within}}^i$$

where $F_{\text{within}}^i$ is the $F_{\text{within}}$ value at $F_{\text{complexity}} = i$. Note that as ELSA finds new and better solutions (with higher $F_{\text{within}}$), the coverage increases. Note also that when ELSA with model $k$ did not explore a single solution at a specific $F_{\text{complexity}}$, its coverage at that $F_{\text{complexity}}$ value is set to zero. Models with higher coverage are regarded as better models than models with lower coverage.

![Candidate Fronts](image)

**Figure 3. Candidate fronts of three models in Experiment 1**

Table 1 shows the means (average of coverages defined above) and standard deviations of three models over 15 runs. The $p(i, j)$ represents a $t$-test $p$ value that compares the means of Model $i$ and Model $j$. For comparison purposes, the means of three models are computed over two different regions: one with all $F_{\text{complexity}}$ values and the other with $F_{\text{complexity}} > 0.66$. Table 1 shows that the means of Model 2 and Model 3 are significantly different from that of Model 1 in the region with $F_{\text{complexity}} > 0.66$. However, the three models are not significantly different in the region with all $F_{\text{complexity}}$ values.

<table>
<thead>
<tr>
<th>Region</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All $F_{\text{complexity}}$</strong></td>
<td><strong>Coverage</strong></td>
<td><strong>Coverage</strong></td>
<td><strong>Coverage</strong></td>
</tr>
<tr>
<td></td>
<td>21.4136 ± 0.3875</td>
<td>21.6321 ± 0.4823</td>
<td>21.4003 ± 0.4219</td>
</tr>
<tr>
<td>$F_{\text{complexity}} &gt; 0.66$</td>
<td><strong>Coverage</strong></td>
<td><strong>Coverage</strong></td>
<td><strong>Coverage</strong></td>
</tr>
</tbody>
</table>

**Table 1. The coverage of three models in Experiment 1: Model 1 (model without relevance metric), Model 2 (model with boolean value metric) and Model 3 (model with real value metric)**


Experiment 2

Data Description

In Experiment 2, the effects of two new evaluation metrics are tested on the real datasets taken from the CoIL 2000 forecasting competition. For more information about the CoIL competition and the CoIL datasets, refer to the Web site http://www.dcs.napier.ac.uk/coil/challenge/. In this study, the evaluation data with 4000 household records is used and each record contains 93 variables, containing information on both socio-demographic characteristics (variables 1-51) and ownership of various types of insurance policies (variables 52-93). Details are provided in Table 2.

<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of houses owned by residents</td>
</tr>
<tr>
<td>2</td>
<td>Average size of households</td>
</tr>
<tr>
<td>3</td>
<td>Average age of residents</td>
</tr>
<tr>
<td>4-13</td>
<td>Psychographic segment: successful hedonists, driven growers, average family, career loners, living well, cruising seniors, retired and religious, family with grown ups, conservative families, or farmers</td>
</tr>
<tr>
<td>14-17</td>
<td>Proportion of residents with Catholic, Protestant, others and no religion</td>
</tr>
<tr>
<td>18-21</td>
<td>Proportion of residents of married, living together, other relation, and singles</td>
</tr>
<tr>
<td>22-23</td>
<td>Proportion of households without children and with children</td>
</tr>
<tr>
<td>24-26</td>
<td>Proportion of residents with high, medium, and lower education level</td>
</tr>
<tr>
<td>27</td>
<td>Proportion of residents in high status</td>
</tr>
<tr>
<td>28-32</td>
<td>Proportion of residents who are entrepreneur, farmer, middle management, skilled laborers, and unskilled laborers</td>
</tr>
<tr>
<td>33-37</td>
<td>Proportion of residents in social class A, B1, B2, C, and D</td>
</tr>
<tr>
<td>38-39</td>
<td>Proportion of residents who rented home and owned home</td>
</tr>
<tr>
<td>40-42</td>
<td>Proportion of residents who have 1, 2, and no car</td>
</tr>
<tr>
<td>43-44</td>
<td>Proportion of residents with national and private health service</td>
</tr>
<tr>
<td>45-50</td>
<td>Proportion of residents whose income level is $&lt;30,000, $30,000-$45,000, $45,000-$75,000, $75,000-$123,000, &gt;$123,000, and average</td>
</tr>
<tr>
<td>51</td>
<td>Proportion of residents in purchasing power class</td>
</tr>
<tr>
<td>52-72</td>
<td>Scaled contribution to various types of insurance policies such as private third party, third party firms, third party agriculture, car, van, motorcycle/scooter, truck, trailer, tractor, agricultural M/C, moped, life, private accident, family accidents, disability, fire, surfboard, boat, bicycle, property, social security</td>
</tr>
<tr>
<td>73-93</td>
<td>Scaled number of households holding insurance policies for the same categories as in scaled contribution attributes</td>
</tr>
</tbody>
</table>

Table 2. Household background characteristics

Experimental Results

As in Experiment 1, three models are compared in Experiment 2 with few modifications. In model 2, 0 is assigned for the first 51 features and 1 for the remaining features as the relevance value based on a well known marketing science work (Rossi et al., 1996). In model 3, 0.25 is assigned for the first 21 features, 0.5 for the next 30 features (attributes 22-51), and 0.75 for the last features (attributes 52-93) as relevance values.

Table 3 numerically summarizes coverages of three models on a real data set. Because of computational complexity due to the high dimensionality of input variables and number of records, three models are evaluated only once in Experiment 2. In experiment 2, the region with $F_{\text{complexity}} > 0.50$ are particular interests because when all significant 42 features (attribute 52-93) are used for clustering, its $F_{\text{complexity}}$ becomes approximately $1 - (42-1)/(93-1) \approx 0.5543$. In Table 3, the model with a real value relevance metric (Model 3) shows the best performance both over high $F_{\text{complexity}}$ values and all $F_{\text{complexity}}$ values. In contrast, clustering model with boolean value relevance metric (Model 2) shows comparable performance over two area. Therefore, in terms of coverage and robustness on synthetic and real data sets, real value relevance metric is recommended for clustering purpose.

Proceedings of the Tenth Americas Conference on Information Systems, New York, New York, August 2004
Table 3. The coverage of three models in Experiment 2: Model 1 (model without relevance metric), Model 2 (model with boolean value metric) and Model 3 (model with real value metric)

<table>
<thead>
<tr>
<th>Region</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>All $F_{complexity}$</td>
<td>48.7281</td>
<td>49.5823</td>
<td>54.5394</td>
</tr>
<tr>
<td>$F_{complexity} &gt; 0.55$</td>
<td>24.7284</td>
<td>24.6394</td>
<td>29.2653</td>
</tr>
</tbody>
</table>

Visualization of Candidate Fronts

This paper also introduces a simple but effective visualization tool to help decision makers better understand each candidate solution and make a choice out of available solutions. This tool takes all evaluated solutions by ELSA as an input and constructs candidate fronts with additional information including the values of all evaluation criteria of clusterings. The screen shots of outputs of this visualization tool are shown in Figure 4.

![Figure 4](image-url)

Figure 4. Screen shots of candidate fronts constructed using all evaluated solutions from a model with a real value relevance map in Experiment 2

The left panel of the output shows candidate solution ID, the number of clusters constructed, and the values of evaluation metrics. In the right panel, it draws the candidate front with $F_{within}$ values of all candidate front solutions. The most important feature of this tool is that it allows the decision maker to move along the candidate front using an arrow key on the keyboard. Another way to move the candidate front is to click on one of $F_{within}$ values in the vertical scrollbar. As the decision maker uses the up-arrow key to examine the solutions on the candidate front, a little marker indicates the current position. The values of evaluation metrics of the corresponding solution are automatically updated in the left panel. This way the decision
maker can easily identify few interesting solutions that satisfy minimum criteria (e.g., $F_{within} > 0.9$ and $F_{interest} > 0.5$). Assume that four solutions in Table 4 are selected by the decision maker. Among four solutions, Solution 59 can be easily removed from further consideration because other solutions are superior over most criteria. However, it is difficult for the decision maker to choose one of the remaining solutions because neither one dominates others over all criteria. For example, solution 62, Solution 64, and Solution 68 are the best in terms of $F_{within}$, $F_{relevance}$, and $F_{complexity}$ respectively. Therefore, the final decision will be made based on the decision maker’s subjective weight on these criteria and other organizational and procedural constraints. By making it possible to easily identify “good” solutions for further consideration, the presented visual tool can greatly help decision makers make a final decision.

<table>
<thead>
<tr>
<th>Solutions</th>
<th>$F_{complexity}$</th>
<th>$F_{within}$</th>
<th>$F_{relevance}$</th>
<th>$F_{interest}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solution 59</td>
<td>0.7500</td>
<td>0.9146</td>
<td>0.6354</td>
<td>0.5593</td>
</tr>
<tr>
<td>Solution 62</td>
<td>0.7826</td>
<td>0.9405</td>
<td>0.6190</td>
<td>0.5809</td>
</tr>
<tr>
<td>Solution 64</td>
<td>0.8043</td>
<td>0.9403</td>
<td><strong>0.6842</strong></td>
<td>0.5583</td>
</tr>
<tr>
<td>Solution 68</td>
<td><strong>0.8478</strong></td>
<td>0.9278</td>
<td>0.6500</td>
<td>0.5623</td>
</tr>
</tbody>
</table>

Table 4. Solutions identified from the visual tool for further consideration

CONCLUSION

This paper introduces two human-driven metrics to evaluate the quality of clusterings with two traditional data-driven metrics. For comparison purposes, three different models are used to incorporate the relevance metric and compared on a synthetic and a real data sets. The key findings are summarized as follows.

- Two new measurement metrics—$F_{relevance}$ and $F_{interest}$—are successfully incorporated into an evolutionary algorithm to reflect decision makers’ subject confidence in the importance of each feature to the clustering and decision making process.
- An evolutionary algorithm optimizes the multiple criteria separately while exploring more promising space of possible feature combinations with relevance maps specified by decision makers. The presented model allows human experts to input their expertise into decision support systems while minimizing the interaction.
- The model with a real value relevance map returns the highest coverage of candidate fronts and allows human experts to input their true confidence as close as possible into the system. In contrast, the model with a boolean value relevance map only allows boolean types of confidence.
- Most importantly, through the proposed graphical interface, decision makers choose few candidate fronts for further detailed analysis after considering all criteria.

In future work the effects and significance of two new metrics on the quality of clusterings will be tested in various domains of applications. In particular, estimating individual effect of each criteria and finding any relationship among criteria will be critical to better understand and design a new decision support system. Another direction of future work is to develop a more elegant graphical interface. The most important component of the new graphical interface will be the connection between each of candidate solutions and summary information of clusterings including the number of records in each cluster, the centroids of used variables, and so on. This way decision makers can have a tool that supports both analysis and implementation at the same time.

REFERENCES


