Editor: Mike Potel
www.wildcrest.com

Visual Classification: Expert Knowledge Guides Machine Learning

Joseph MacInnes, Stephanie Santosa, and William Wright, ***Oculus Info***

Domain experts can effectively guide the classification of the activities by visualizing user activity as a two-dimensional space. This visual classification method is presented as part of a four-phase analytic workflow model detection process.

Classes or labels are one type of information we can extract from data sets. Classifying and labeling something not only gives us information on what it is but also provides knowledge about how it relates to other things we observe. Humans have evolved to be natural and accurate classifiers, but our ability to categorize objects or create new categories is occasionally limited not by our ability to classify but by our ability to perceive the attributes and patterns needed to form these new classes. Critical classification attributes might be hidden within hundreds of other variables (as with biometrics). Classification patterns that are part of a time series could tax human memory limits (as with medical imaging). In addition, classification problems might need to be solved very quickly or on very large data sets (as with Internet document sorting).

Classification in machine learning tries to solve these problems by automating the classification process. The number of techniques and algorithms is vast and (aptly) can be classified in many ways. All machine-learning classification algorithms learn by example. They begin with a large set of observations belonging to one of at least two classes, with the goal of learning a way to distinguish the class in the current observations and future, unseen observations. Classic problems include automatically sorting science articles from different disciplines, determining whether x-rays contain images of tumors, and parsing sound recordings for words. Also, as the Web moves toward a semantic representation, the need for machines to understand large volumes of information will require classification algorithms with machine speed and human understanding.

We believe that a mixed-initiative approach combining human and machine reasoning can create such algorithms. Here, we show how we've used this approach to classify user interactions with a software application for intelligence analysis into domain-specific analytic activities

Mixed-Initiative Classification

Machine-learning tools have reached a level of maturity in which developers and users can use them without having to be experts in machine learning. Open source tools let users of all skill levels create Bayesian and neural networks with a visual drag-and-drop interface.

New ways of visualizing machine-learning classifiers have allowed some human input into the classification process itself. We want to develop creative visualizations that get these powerful tools into the hands of the domain experts who will be using these classifiers. For example, if we aim to create a workflow model to use as an intelligent calculus tutor, a calculus teacher would best understand how to deliver the material. So, the teacher, not the machine-learning programmer, should guide the workflow discovery.

Kayur Patel and his colleagues[1] suggest that software developers applying stochastic machine learning have difficulty

- following these algorithms' iterative nature,
- understanding the model and how the input relates to the results, and
- evaluating the model's performance.

Of these difficulties, "understanding the model" is probably the most critical for domain experts such as our calculus teacher. Understanding the relation between the observations and the resulting classification not only helps the expert build better models but also creates an appropriate level of trust in the classification, which is critical in any automated system. Furthermore, because many of these algorithms work on processed observations,

 Published by the IEEE Computer Society

Machine Categorization

Machine categorization attempts to group observations into class sets on the basis of observation features or statistical similarity. Algorithms differ in the amount of prior information that's available about the classes.

Supervised algorithms, such as neural networks (NNs), require that the training observations are prelabeled according to the class to which they belong. The prelabeling gives these algorithms a way to test their guesses against the "true" class for that observation. For example, an NN could be given a list of images containing handwritten letters along with the actual letter's labels. The NN will learn to associate the pictures with the actual letter (class) and apply this learning to new images that haven't been labeled.

Unsupervised algorithms, such as *k*-means clustering, don't have labeled observations and must rely on grouping similar observations on the basis of some similarity metric. For example, given the same images as the NN, a successful unsupervised algorithm would place similar images in 26 different bins but wouldn't know what these bins meant.

Semisupervised algorithms try to work with only some of their observations labeled.

Patel and his colleagues suggest there be a way to reference back to the original raw data. For our domain expert, this would be putting that data in a visual context that's easier to understand—that of the learning problem.

Human expert input into the machine-learning process has traditionally been limited to labeling data for supervised learning and setting prior probabilities in Bayesian networks. We propose an approach that integrates domain expert input into the process. This mixed-initiative guidance isn't analogous to supervised learning. The question of supervision (see the "Machine Categorization" sidebar for details) refers to whether individual observations used for classification are labeled as belonging to a particular class. "Supervised" refers to an expert's input before classification, whereas our approach allows expert input during classification. Figure 1 illustrates the involvement of the user—the domain expert—and the machine-learning system in the mixed-initiative classification process toward creating a workflow model.

Allowing human expert input by providing data visualization during machine learning will improve the accuracy of the resulting classifications where possible. More important, it will increase result comprehensibility and trust.

Figure 2 shows our four-phase proposal for detecting analytic workflows in free, unstructured analysis. Analytic workflow is the process of gaining understanding from information that's incomplete, noisy, and often intentionally misleading. Phase 3 involves our mixed-initiative clustering, which supplements the more traditional supervised and unsupervised data segmentation. Later, we show how to visualize our mixed-initiative classification as an initial part of creating complex cognitive workflow models.

Method

As the analytic software environment for our experiment, we chose nSpace.[2] nSpace is an integrated cognitive workspace used in information analysis. It comprises

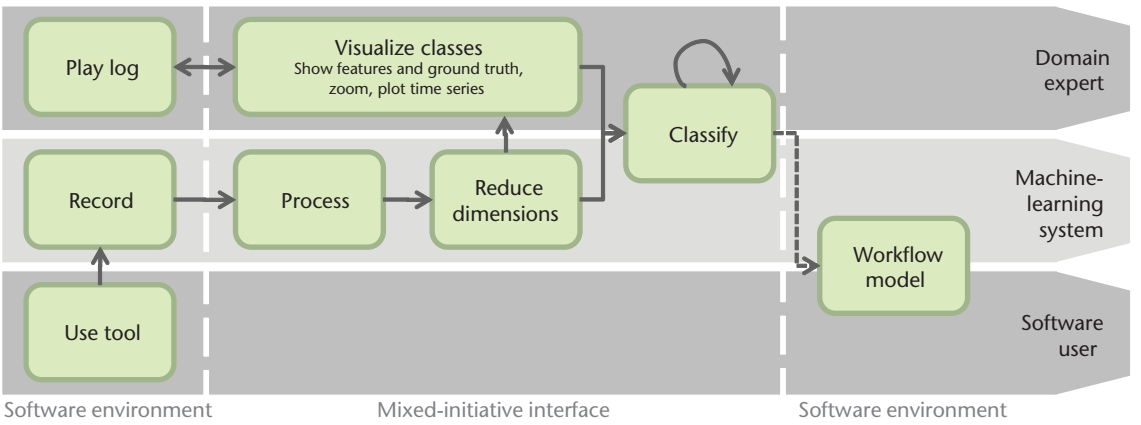

Figure 1. Expert-guided clustering. First, the user records sensor data. A domain expert then uses a visual interface to guide a machine-learning algorithm through a classification process to create a workflow model.

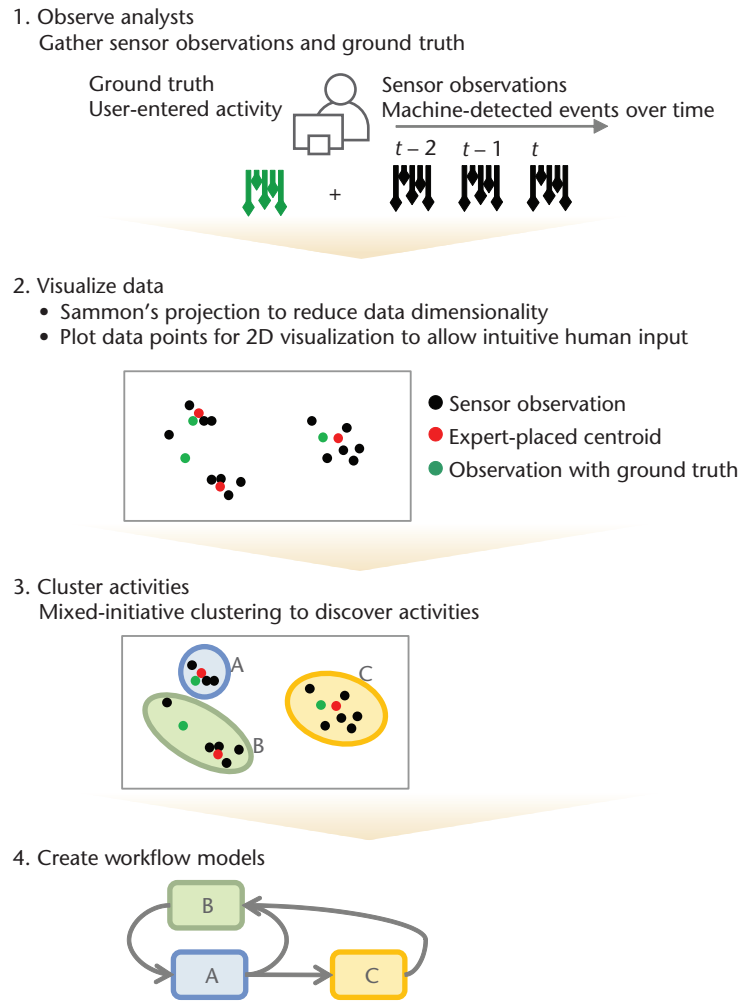

Figure 2. A four-phase process for workflow detection. Domain experts in analytic workflow guide machine-learning algorithms to create more intuitive workflows. For details on machine classification and Sammon's projection, see the related sidebars.

Sammon's Projection

To get human input into our mixed-initiative clustering, our high-dimensional clustering data must be transformed into something people can read. Sammon's projection reduces the data's dimensionality while maintaining the relative interobservation distance. It also reduces this distance by minimizing Sammon's stress between observations. Sammon's stress is defined as

$$E = \frac{1}{\sum_{i<j} d_{ij}^*} \sum \sum_{i<j} \frac{\left(d_{ij}^* - d_{ij}\right)^2}{d_{ij}^*},$$

where d_{ij} is the distance between the ith and jth points in the original high-dimensional space and d_{ij}^* is the interpoint distance in the 2D space. We used the Euclidean distance between points for our tests, although other distance measures are possible. Error was reduced at each iteration through standard gradient descent of Sammon's stress.

- TRIST (The Rapid Information Scanning Tool), for information triage;
- Sandbox, for evidence marshaling and visual sense-making;
- Viewer, for information and entity extraction; and
- Projects, for project planning and collaboration.

These tools complement each other to support a full set of information analysis tasks.

Figure 3 illustrates the nSpace suite of tools and how a workflow could span across them. A user can organize a project in Projects and then query information in TRIST, which presents the relevant documents and entities within them. The user can view selected documents in Viewer and extract snippets and entities from both the Viewer and TRIST to analyze in Sandbox.

Data Collection

We chose nSpace because it supports a wide variety of analytic workflows.[2] A data-collection service can also use TRIST's Web-based architecture to collect observational information. nSpace is equipped to detect *analytic log events* (ALEs)—high-level sensors used to describe events such as modifying evidence and making claims. These ALEs, along with application-specific information, captured analyst behavior during a two-hour task in the nSpace environment. We used these ALEs to discover analytic activities and then our workflow models.

Six participants performed an analysis on the trade relationship between two countries. We intentionally left the task open-ended to allow for more open-ended workflows from the participants. All participants were students who had completed at least one year of a specialty program in incisive analysis. We gave them a two-hour tutorial on nSpace and instructed them to use whatever analytic techniques and workflows they deemed appropriate for the task.

At five-minute intervals during the experiment, the software prompted participants to enter their current analytic activity. The software presented the options from a hierarchy created for this experiment. The hierarchy had seven top-level categories: Plan, Search, Examine, Marshal, Reason, Collaborate, and Report. The "Analytic-Events Hierarchy" sidebar lists all the activities.

If the hierarchy didn't include the participants' current activity, they could enter their own descriptors. These entries served as "ground truth" labels for the corresponding sensor data used during clustering.

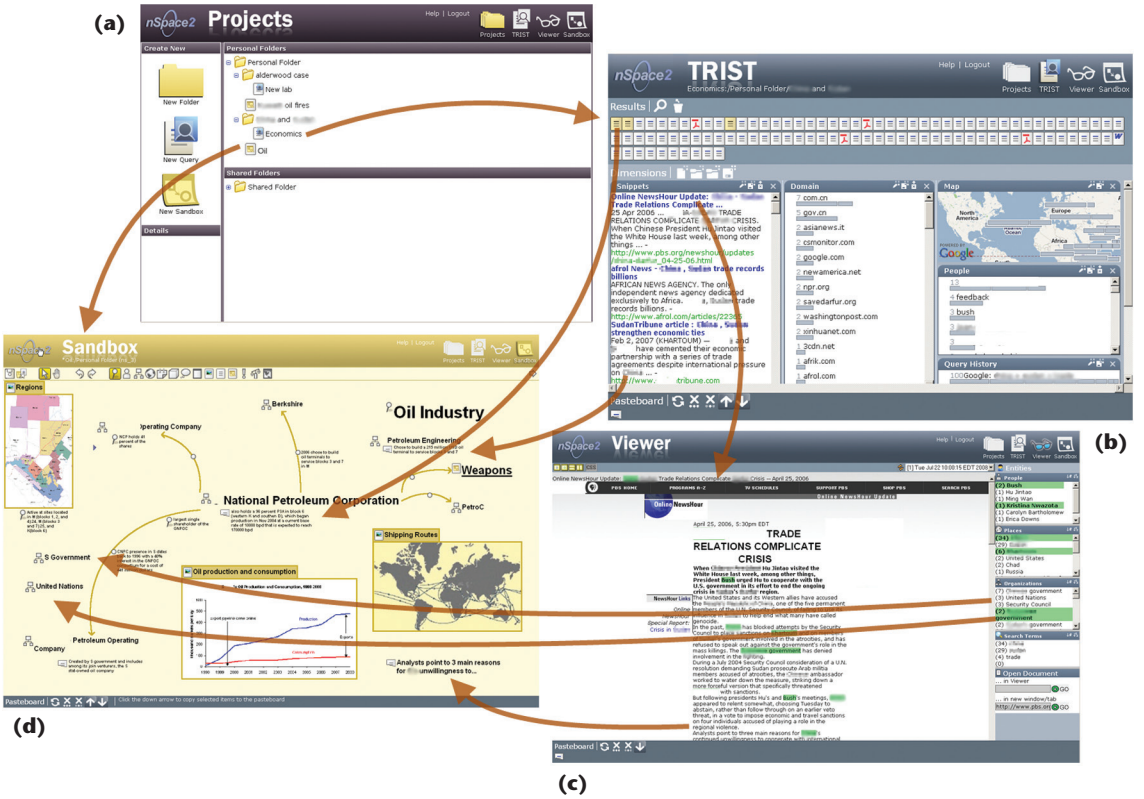

Figure 3. Participants used nSpace, a Web-based analytic environment, to complete their task. nSpace is a system of systems consisting of (a) Projects, for project planning; (b) TRIST (The Rapid Information Scanning Tool), for information triage; (c) Viewer, for information and entity extraction; and (d) Sandbox, for analytic sense-making. They combine to support complete analytic workflows, and information is easily transferred between them.

Mixed-Initiative Clustering

We preprocessed the data from the experiment to achieve the input requirements for Sammon's projection[3] (see the related sidebar) and mixed-initiative clustering. This involved first normalizing the data from each sensor according to the

Analytic-Events Hierarchy

We presented this hierarchy to participants to extract ground truth during the experiment. It also served as labels during the mixed-initiative clustering and will become hidden states in phase four of dynamic Bayesian networks.

1. **Plan**
   a. Make To-Do List
   b. Group Tasks
   c. Name New Task
   d. Configure Tool
   e. Switch Tool/Method
   f. Version
2. **Search**
   a. Query
   b. Monitor
   c. Query Refinement
   d. Scan
   e. Filter
3. **Examine**
   a. Read/Listen/Watch
   b. Compare
   c. Correlate
4. **Marshal**
   a. Arrange/Move
   b. Create Categories
   c. Categorize
   d. Extract
   e. Merge
5. **Reason**
   a. Create Hypothesis
   b. Construct Argument
   c. Change Perspective
   d. Mark Leverage Points
   e. Rearrange
   f. Link
   g. Challenge
   h. Make Notes
   i. Annotate Information
   j. Assess
6. **Collaborate**
   a. Asynchronous Review
   b. Informal Synchronous Review
   c. Make Presentation
7. **Report**
   a. Compose
   b. Edit
   c. Define Milestones
   d. Summarize

range observed across all participants. This project's eventual goal is to extract workflow patterns that fit across all analysts. So, normalization over the complete, observed range is required for comparison.

To ensure that no two observations were identical, we then used recursive randomization to slightly perturb observations within a subject containing identical sensor values. Because "sammonizing" is based on the interobservation distance, the data must have no zero distances. Duplicate removal or various windowing techniques are possible at this stage; we're considering them for future research when more data is available.

We designed the visualizations and interactions in our mixed-initiative interface to make activity patterns visible to the domain expert and to allow exploration and marking of those patterns. The expert first sees any apparent clusters after Sammon's projection, along with the ground-truth labels the user provides in Sandbox. For our purposes, partial labeling (see the last paragraph of the "Machine Categorization" sidebar) of the data using ground-truth labels not only is less work than fully labeled data but also makes for a cleaner interface. We do, however, run the risk that a key cluster won't be labeled; however, other exploration tools will let the expert make a reasonable guess at its activity label.

Mixed-initiative machine-learning techniques enable the production of machine-learned models that end users can trust and find useful.

Unlike many AI approaches that act like automated processes, our guided clustering allows domain expert input throughout the process. Experts can specify the number of clusters and how many training iterations they require. They can set the initial center position of all clusters and choose the most appropriate label. Data visualization in the Sammon-projected space lets the expert gain enough understanding of the data to provide these inputs. Setting the initial activity cluster centroid position to an intuitive location will likely improve performance over standard *k*-means clustering, which uses random locations. Our algorithm then incorporates the expert's suggestions for cluster centroid parameters with the *k*-means' suggestions for cluster centroids to combine both expert intuition and statistical accuracy. Experts can also override the *k*-means input for a single cluster centroid by using the Force Centroid option or for all centroids by disabling *k*-means completely.

Visualization of activity clusters over time is critical for this process because the observations are time-based. Although the real temporal modeling occurs after these activity clusters are formed (see phase 4 in Figure 2), letting our experts view observations over time will provide extra information for the guided clustering (phase 3). For example, by watching a user's data move back and forth over time between two obvious activity clusters, an expert can use ground-truth labels in one of those clusters to infer the second cluster's label. If a user is transitioning between "Create Categories" and an unknown cluster in Sandbox, the expert might label the second cluster "Categorize."

Our visualizations also make feature extraction easier. With some algorithms, it's required or beneficial to restrict the number of features (our ALE sensors in this case) that each observation uses, to reduce the work of clustering. Our interface can provide feedback on which features were most important in determining cluster assignment. Although our software saves dozens of features as ALE sensors during each observation, they won't all be important during all activities. By providing the statistical importance of "importing evidence" to marshaling activities, experts can use this information, in addition to their own intuition, during workflow creation (phase 4 in Figure 2) to reduce observation complexity. Finally, because the software saves observation features as sensor log events and their times, experts can return to the original software package to see all observations in context with the user's task. Patel and his colleagues suggest that this ability is critical to understanding the final model.[1]

Results

Figure 4 displays one participant's sandbox (see Figure 4a) and cluster data (see Figure 4b). Experts can control the number and the clusters' initial centroids, forgoing the need for an algorithm solution to these problems. By allowing the analytic expert choice over centroid initialization, our clustering algorithms should reach an intuitive solution in less time than random initialization.

To produce the clustering's results, we use a hybrid approach that averages a final cluster centroid, which is a weighted average of

- the expert's centroid ($w = 0.4$),
- a *k*-means centroid ($w = 0.5$), and
- a centroid based on the locations of any ground truths ($w = 0.1$).

The weight accorded to each of these inputs is open to future research, which experts could adjust in future versions. We set the ground-truth weight low in this trial owing to our observation that experts tend to rely heavily on these points in setting their own centroids, and a higher weight would overemphasize this data.

As Figure 4 shows, the initial results were promising. Participant MH1's observations in Sandbox would typically present difficulty for both human and automatic clustering. Correlate and Categorize are fairly clear clusters, and our ground-truth entries provide reasonable activity labels. However, the collection of Plan ground-truth entries in the middle top would provide difficulties for both machine and human solutions on their own. A machine solution might not pick this up as an activity cluster because it's quite close to other observations on either side of it. A human would spot these labels but would have difficulty finding the boundary between this activity group and others.

Our approach lets the domain expert signal that an activity cluster is centered here, and the *k*-means component will easily set the boundaries on the basis of the observation distance. Our human domain expert was also able to change an activity label that seemed out of place. Participant MH1 used the Plan/Group Tasks label in this sandbox, even though we intended this to be for planning and grouping daily tasks. On the basis of the current context, our expert decided that the participant meant the "task of grouping evidence" and relabeled the activity cluster as Merge. A pure *k*-means algorithm can estimate cluster labels on the basis of choosing the most common ground-truth label for that cluster, or perhaps the label closest to the cluster centroid. The contextual judgments shown here, however, are possible only in a mixed-initiative algorithm.

Although we haven't yet fully automated creation of Bayesian workflow models using these activities, the "Marshaling by Hypothesis" sidebar demonstrates this research's potential. A domain expert created a workflow for marshaling by hypothesis, which demonstrates the workflow process by which an analyst extracts useful information, given a question or hypothesis. Although the fit isn't yet perfect, the workflow we created using our mixed-initiative categories is much closer to the domain expert's than the one we created using the ground-truth categories alone.

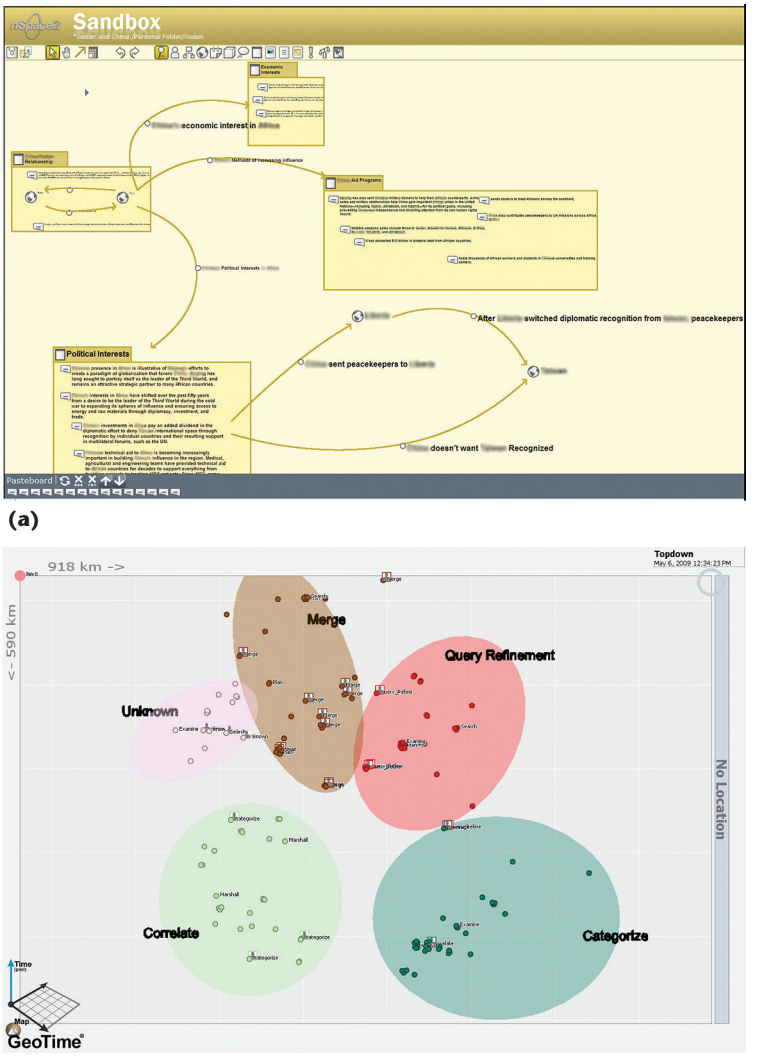

Figure 4. Example of (a) participant MH1's work and (b) the resulting clusters we created via visual clustering. To achieve these results, the domain expert contributed the number of clusters, the initial centroid locations, and iterative adjustments to the *k*-means algorithm.

Mixed-initiative machine-learning techniques enable the production of machine-learned models that end users can trust and find useful. Traditional black-box solutions might have been accurate but often couldn't explain why a recommended solution was correct. Mixed-initiative solutions increase transparency and understanding by

- utilizing accurate visualizations of patterns to allow for human expert input to the process,
- providing contextual knowledge that the AI might not be aware of,

Marshaling by Hypothesis

We had a human expert specify a workflow for marshaling by hypothesis (see Figure A1). This iterative, nonlinear workflow is modeled as a dynamic Bayesian network and required knowledge of Bayesian probability and visual analytics. We would like our mixed-initiative process (see the main article) to learn this workflow automatically from observations. Figure A2 shows a machine-detected workflow we created using only ground-truth labels from the experiment described in the main article; Figure A3 includes extra information learned from mixed-initiative activities.

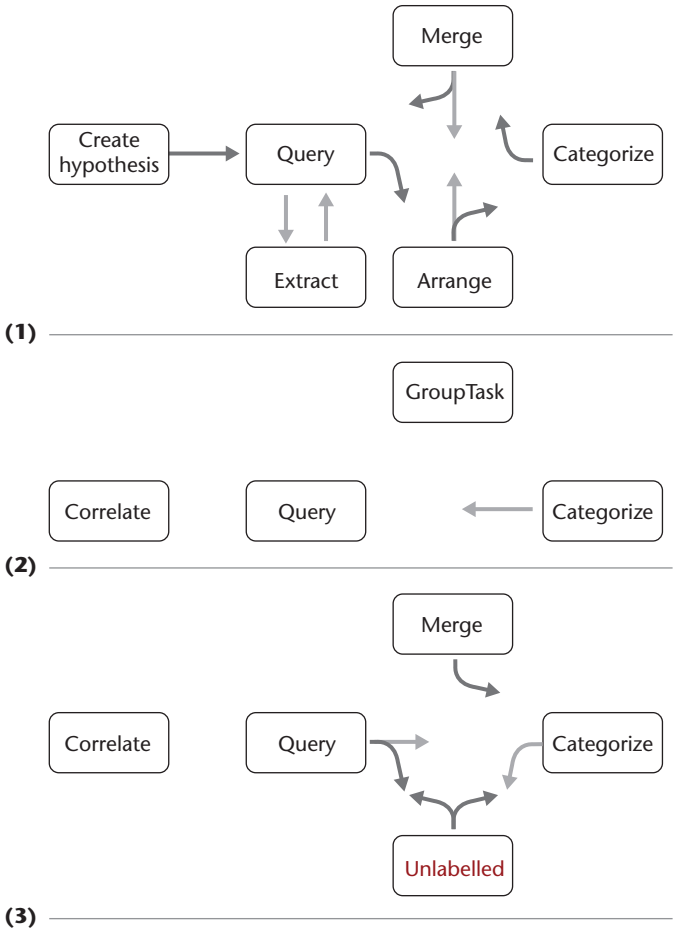

Figure A. Workflows for marshaling by hypothesis. (1) The target workflow, which was created by a human expert, (2) a machine-detected workflow based on ground truth only, and (3) a machine-detected workflow based on ground truth plus mixed-initiative activities. By including mixed-initiative activities, we can discover workflows that match the target workflow more closely than ground-truth activities alone.

- providing human-understandable descriptions and labels of machine-produced patterns and clusters, and
- keeping the expert modeler's intuition in the loop during pattern detection.

Future research will include training separate activity models, likely naïve Bayesian networks, using the clusters we discovered in the mixed-initiative process. We'll then use each model as part of an ensemble method to detect activity states in a dynamic Bayesian network (DBN) of the workflow process itself. Because we want to create clear, accurate workflow models, we'll measure both clarity and accuracy to determine DBN suitability and to check for any inherent trade-off in these two measures. Initial workflows created using this technique (see the "Marshaling by Hypothesis" sidebar) demonstrate potential for our mixed-initiative process, although a full qualitative analysis will require more data.

Although our clustering process is guided and completely interactive, our Sammon projection isn't yet. Owing to our data's high-dimensional nature, it's difficult to begin expert visualizations until the projection is complete. We plan to look into this to expand expert input into clustering. Finally, in our experiments, we played the part of the domain experts. We plan to conduct usability testing on the interface to make the process intuitive to domain experts without a machine-learning background.

References

1. K. Patel et al., "Examining Difficulties Software Developers Encounter in the Adoption of Statistical Machine Learning," *Proc. 23rd AAAI Conf. Artificial Intelligence* (AAAI 08), AAAI Press, 2008, pp. 1563–1566.
2. P. Proulx et al., "nSpace and GeoTime: A VAST 2006 Case Study," *IEEE Computer Graphics and Applications*, vol. 27, no. 5, 2007, pp. 46–56.
3. J.W. Sammon, "A Nonlinear Mapping for Data Structure Analysis," *IEEE Trans. Computers*, vol. 18, no. 5, 1969, pp. 401–409.

***Joseph MacInnes** is a researcher at Oculus Info. Contact him at joe.macinnes@oculusinfo.com.*

***Stephanie Santosa** is a visualization developer at Oculus Info. Contact her at ssantosa@oculusinfo.com.*

***William Wright** is a senior partner at Oculus Info. Contact him at bill.wright@oculusinfo.com.*

Contact Applications department editor Mike Potel at potel@wildcrest.com.