CEDARS: COMBINED EXPLORATORY DATA ANALYSIS RECOMMENDER SYSTEM

Mark A. Livingston¹, Stephen Russell², Jonathan W. Decker¹, and Antonio Gilliam³

¹Naval Research Laboratory
²Army Research Laboratory
³Strategic Analysis, Inc.
Goal

- Capture a domain expert’s approaches for data analysis
- Be able to intelligently recommend or automatically apply these approaches to future analyses (by the same or other analysts)
- Automate analysis of complex data sets
- Help novice analysts increase their expertise
- Assist domain experts in creative exploratory analysis
- Unify architectures for EDA with systems to automate layout and (ultimately) visual representation
Why use Recommender System (RS)?

- Data analyst’s questions: what data should I explore? what analytics should I apply?
  - Transform: ‘What items are relevant?’ ‘What services complement those items?’

- Selecting analytic operations can be cumbersome

- Analyst may overlook appropriate operations due to familiarity bias

- Enhance creativity under ambiguity and uncertainty, which is often an element of exploratory data analysis (EDA)
Why use RS for EDA?

- Confirmatory analysis is “easy to computerize” [Tukey]
- Common tasks where RS provide benefit [Herlocker et al.]
  - Find some good items
  - Annotation in context (emphasize items based on user preference)
  - Recommend a sequence
  - Recommend a bundle
  - Help with browsing
  - Improve the profile by integrating user preference into the decision making task
### Previous Work

<table>
<thead>
<tr>
<th>Adaptive EDA</th>
<th>RS in Workflows</th>
</tr>
</thead>
</table>
| ForceSPIRE [Endert et al.]
- Adjusts layout by changing weights via capturing semantics of user interaction | Optimize hyperparameters
- Improved prediction in retail applications [Chan et al.]
- Improved recommendations by combining machine learning with rules [Bergstra&Bengio] |
| [Petasis et al.]
- Use C4.5 decision tree algorithm to discover need to update rules in recognition and classification of named entities in text corpora | Inference & logic
- Contextually-aware RS [Adomavicius& Jannach] |

CEDARS attempts to bridge gap between adaptive EDA and RS in workflows.
System Architecture

**Layer 1:**
Virtualization (VirtualBox)

**Layer 2:**
- Virtual Machine (Ubuntu)
- Virtual Machine (Ubuntu)
- Virtual Machine (Ubuntu)

**Layer 3:**
- Agent
- Agent
- Agent
- R (R Server)
- Agent
- Agent
- Agent
- R (R Server)
- Agent
- Agent
- Agent
- R (R Server)

**Layer 0:**
Physical Hardware
Django web framework with Python scripts to ingest data
Stores data in MongoDB
Passes interest values (recommendation) to agents
Agents use R for statistical computation
EDA layer collects data from processing agents as plain text, parsed and loaded into Django
User interface accesses local Django server with web browser
Use Case 1

- Study of image metrics on multivariate visualizations
  - Do the image metrics offer insight into user performance?
  - Approximately 600 measures

- Recommendations
  - Summary statistics
  - Requested F-tests
  - Non-requested F-tests
    - visualization technique on edge strength (Sobel) showed difference (target versus distractions)
Use Case 2

- Data from eye tracking, mostly unexplored
- Initial recommendations are for summaries of variables
  - Similarity in the distributions of two variables led to the discovery of data error
- Concern with pupil diameter measurements led to summaries, correlations, and repeated-measures ANOVAs involving those variables
  - Helped identify a need for more restrictive outlier removal threshold
Series of five human participant studies; goal was to explore connections between the analysis (workflow) for data sets.

First data set: cold start, so defaults to summary statistics.
- User selects dependent variables of interest.
- CEDARS displays group means; some are of interest.
- CEDARS follows with ANOVA, then t-tests (independent variables).

Second set: much the same with better ranking.
- User selects dependent variables, gets group means by selected variables, and user selects results of interest.
- Invokes some rules on the first data set where variables names are the same and leads to new recommendations.
- CEDARS invoked some rules using SubjectID, and user sees that one subject was error-prone and fast.
Third set: Summary operations, group means, ANOVA
- Not much of interest found

Fourth set: Summary operations, group means, ANOVA
- New variable is explicitly requested through summary statistics

Fifth set: Summary operations, group means, ANOVA
- Two new variables requested through summary statistics
- Reclassified from numeric to factor (a standard operation in R)
- CEDARS begins to recommend multi-factor ANOVA operations
- CEDARS applies type change to variables with same name in fourth data set
CEDARS can replicate standard analytical practice and provide deep analysis by recommending operations on variables a domain expert had not thought to test.

CEDARS can replicate analysis applied to one data set to another with similar structure or shared names:
- Can be invoked “forward” on new data or “backward” on data in memory.

Ultimate goal of EDA: tell the story that explains the data.

CEDARS can potentially:
- Capture expertise of domain expert and data scientist.
- Use that expertise to guide novices.
- Remind experts of forgotten analytical options.
- Promote adoption of novel analysis methods.
- Unify architectures for automating layout or visual representation.

Future: more data and evaluate recommendations across data sets.
Thank you!

- CEDARS: Combined Exploratory Data Analysis Recommender System technical report (forthcoming)

- Mark.Livingston@nrl.navy.mil