Distil: A Mixed-Initiative Model Discovery System for Subject Matter Experts (Demo)

Scott Langevin SLANGEVIN@UNCHARTED.SOFTWARE
David Jonker DJONKER@UNCHARTED.SOFTWARE
Christopher Bethune CBETHUNE@UNCHARTED.SOFTWARE
Uncharted Software Inc., Toronto, ON, Canada

Glen Coppersmith GLEN@QNTFY.COM
Casey Hilland CASEY@QNTFY.COM
Qntfy, Arlington, VA, USA

Jonathon Morgan JONATHON@NEWKNOWLEDGE.IO
Paul Azunre PAUL@NEWKNOWLEDGE.IO
New Knowledge, Austin, TX, USA

Justin Gawrilow JUSTIN.GAWRILOW@GMAIL.COM
Jataware, Arlington, VA, USA

Abstract

We present in-progress work on Distil, a mixed-initiative system to enable non-experts with subject matter expertise to generate data-driven models using an interactive analytic question first workflow. Our approach incorporates data discovery, enrichment, analytic model recommendation, and automated visualization to understand data and models.

Keywords: Augmented Intelligence, Machine Learning, Mixed-initiative, Visual Analytics

1. Introduction

Few tools exist to make data analysis accessible and conversational. Such tools lack support for the principled top-down, questions-first, approach needed to put the power of machine learning and data science into the hands of non-experts with key subject matter knowledge. Because many data science tasks are procedural, there is an opportunity to (semi-)automate them, closing the loop between subject matter experts (SMEs) and decision-making.

Our approach to addressing this challenge is *Distil*, which uses a question-driven, mixed-initiative approach to maximize the combinatorial power of human/machine intelligence for data-driven discovery of models. Current data analysis tools assume that users are: 1) familiar with the data; 2) knowledgeable about visualization options and limitations; 3) able to pre-process data, extract salient information, and apply analytics to answer analytical questions; and 4) able to "slice and dice" their way to the answer using basic charts and dots on a map. They provide no workflow support for analytic goals, limited assistance or suggestions for model development, understanding, or analytic thinking and limited support for iterative analysis and sensemaking. Our research focus in the development of Distil aims to overcome these limitations by combining semantic data discovery services, analytic

model recommendation, and automated visual analytic data and model summarization to construct quantitative models in support of analytical needs.

In this paper we present in-progress work on Distil and its technical components.

2. Related Work

Previous research has focused on providing assistive agents for various aspects of the data science process. *Data wrangling* constitutes the bulk of the data science process (Kandel et al., 2012), encompassing data parsing, normalizing, cleaning, and imputing missing data attributes. The mixed-initiative approaches of low-level tools such as Data Wrangler (Guo et al., 2011), OpenRefine (OpenRefine, 2018), and commercial systems like Trifacta (Heer et al., 2015) are designed for expert data scientists. Karma (Szekely et al., 2011) assists semi-automated semantic schema inference to map raw data to a normalized schema. Limited to no support is provided to assist with data discovery or mapping data to user objectives.

Mixed-initiative data analysis combines domain expert knowledge and observational reasoning with the bias-free processing of large volumes of data afforded by machine learning techniques. Cooperative generation of decision trees (Ankerst et al., 1999) and the RESIN system for predictive analytics (Yue et al., 2010) are examples of *human-in-the loop* systems that attempt to harness domain expertise by providing intuitive interfaces for direct model parameter adjustment. This can be challenging for users without a data science background, as it requires understanding of the learning algorithms used. ForceSPIRE (Endert et al., 2012), ScatterGather (Hossain et al., 2012), Bixplorer (Fiaux et al., 2013) and Active Data Environment (Cook et al., 2015) are examples of *human-is-the-loop* tools (Endert et al., 2014). These provide familiar interface elements to support task-level analytic reasoning, and transparently translate user interactions into changes to the underlying computational models. While this approach better employs the cognitive ability of domain expert users, establishing the mapping between interface actions and the model can be difficult.

Mixed-initiative analysis can be complemented by natural language interfaces such as those found in DataTone (Gao et al., 2015), Articulate (Sun et al., 2010), and Watson Analytics (IBM, 2018). They provide a way to specify simple goals and manage ambiguity in natural language queries, but are limited to basic charts, summary statistics on a single dataset, and "template" analysis with limited data or analytic recommendations.

While the tools and techniques outlined provide significant automation of data science tasks, they suffer in aggregate as each is tailored to a specific class of model or problem space. Integration of these types of domain-specific approaches into a general analytic framework remains an open problem (Wang et al., 2016) (Makonin et al., 2016).

3. Technical Approach

Distil is a mixed-initiative decision-driven modeling workbench that aims to enable SMEs to discover underlying dynamics of complex systems without the need for rare expertise in data science. Through Distil, SMEs visually explore and understand heterogeneous data sources related to analytic objectives, express the objectives using an intuitive visual vocabulary, and interact with, understand, curate, and refine resultant machine-inferred data models.

The vocabulary of current data science tools consists of computing and plotting variables. Assembly of low-level variables, analytic derivatives, and visualizations *for a purpose* requires expertise that few have, and is far too laborious and fault prone. These approaches often only answer low-level statistical questions in isolation and cannot inform courses of action without expert interpretation, consolidation, and extrapolation.

In contrast to the existing approach of low-level statistical analysis or laborious, obscure and error prone processing pipeline building, we focus on visual question decomposition into quantifiable facets that recommender services compose into user-tailorable analytic workflows by interfacing with model construction components. A mixed-initiative human-computer dialog guides model assembly and refinement with expert knowledge. Our technical approach encompasses the following components:

Data Enrichment and Discovery primitives: A collection of primitives to extract semantic information, identify explanatory relationships and conceptual data descriptions, and characterize analytic utility of datasets for recommending data and tailoring model discovery for analytic goals;

Analytic Model Recommendation Engine: Semi-automated to match data, user analytic goals, and analytic primitives to generate empirical models;

Automated, Adaptive Visual Analytic Recommendation Engine: Semantics-driven to guide user understanding of data, complex models, and generation of tailorable visual analytic workflows for sense-making of model output;

Mixed-Initiative Decision-Driven Modeling Workbench: To express analytic goals using a visual and natural language vocabulary. The Workbench helps users visually understand relevant data, explore and refine models.

3.1. Data Enrichment and Data Discovery

To extract knowledge from data, it is essential to understand and prepare the content for consumption by other components in a model discovery system. Four primary components enrich the data by inferring what it contains and how it might be used: *Novelty Detection*, *Semantic Data Type Classification*, *Concept Mapping*, and *Analytic Data Characterization*.

3.1.1. Novelty Detection

While users sometimes examine data to evaluate a specific, well-formed hypothesis, often they simply explore it to identify available insights. Given the size of possible model spaces, it helps—even for small datasets—to identify potential starting points for exploration.

This is a difficult problem for an automated system, but techniques often associated with the feature selection task can identify starting points for an investigation. Principal Component Analysis (PCA) (Hoffmann, 2007) allows features to be ranked based on their contribution to the variance of the dataset as a whole, while Random Forests (Breiman, 2001) allow for the importance of features relative to a specific target feature to be captured.

In our system, PUNK (New Knowledge, 2017) applies a PCA-based ranking algorithm, where feature weightings of the first principal component or the average weighting across the top N principal components derive feature interestingness. Feature importance ranking relative to a given target is computed using a Random Forest combined with a grid search.

3.1.2. Semantic Data Type Classification

Systematic data labeling attempts with semantic web models like the Resource Description Framework (RDF) have largely focused on domain-specific, supervised dataset categorization for information retrieval (Ben-David et al., 2010). For synthesis, model building, and other analytical tasks, data type classification is still determined manually or by heuristics. This may suit a bespoke analysis, but any system to automate the analysis process must also automate this intuition. A model for learned Semantic Data Type Classification enables automated data enrichment, data synthesis, model construction, and model explanation.

Semantic typing in Distil uses SIMON (Semantic Inference for the Modelling of Ontologies), a Character-Level Convolutional Neural Network model for text classification (New Knowledge, 2018). Given a set of input strings, SIMON computes a list of possible semantic types for the set, along with their associated probabilities. Basic types, such as integer, floating point, and categorical can be inferred, along with richer types such as address, date, and geographic position. By relying on a model of structural features, SIMON avoids the fragility associated with typing based on ad-hoc, human-coded rules.

3.1.3. Concept Mapping

Extending Semantic Data Type Classification, we learn higher-level concepts to understand what is in the dataset and what the data is about in real-world context. We draw on unsupervised learning techniques for textual topic analysis such as Latent Dirichlet Allocation (Blei et al., 2003). We augment these techniques with metadata such as semantic data types and word sense disambiguation with technologies such as WordNet (Fellbaum, 1998), FrameNet (Johnson et al., 2002) and Word2Vec (Mikolov et al., 2013).

Our system DUKE (Dataset Understanding via Knowledge-base Embeddings) (Azunre et al., 2018) employs a pre-trained Knowledge Base semantic embedding to perform type recommendation within a prespecified ontology. We aggregate the recommended types into a small collection of super types predicted to be descriptive of the dataset by exploiting the hierarchical structure of the various types in the ontology.

3.1.4. Analytic Data Characterization

Analytic Data Characterization aims to characterize the suitability of datasets for types of analysis. For example, dataset X contains temporal properties, but the primitive detects that there are no trends, anomalies, or seasonality in the time series, so little is to be gained temporally. Similarly, through extracted relationships we can construct any number of graphs, but are any of them useful? With an ontology to describe the inherent analytical utility of the data, we can inform data recommendation services and guide model discovery according to user-expressed objectives. Further, related data and analytics need a measurable scoring system for determining analytic suitability and quantifiable metrics.

Example scoring opportunities based on data type in terms of utility include: for **time series**, Distribution across binning options, correlation, clustering, peak, trend, and change detection; for **graphs**, modularity, community detection, centrality, connected components, and triangle counting; for **categorical values**, value-counting, entropy scores, and Google-search prevalence; and for **geospatial**, geographic distribution and distance-to metrics.

3.2. Model Query and Analytic Workflow Recommendation Engine

Our model recommendation engine consists one components to continually search the solution space, exploring processing primitive and data interactions (Meta-Learning), and another to use the exploratory results to respond to user queries (Task Learning).

Meta-Learning: This includes a system for continual, automatic, empirical analysis of data science primitives, their interactions with each other, and their interactions with data. These findings are necessary to understand the space of reasonable configurations of primitives, optimal pipelines of primitives, and cohesive pipeline fragments from sequentially ordered primitives. This process continually runs and explores the solution space, populating a repository with results of the experiments. This repository, which is accessible to the processes involved in fulfilling SME queries, can significantly reduce the necessary search space during Task Learning.

Task Learning: The application of knowledge derived from Meta-learning in response to an SME query with a particular set of data. This is a general framework for drawing from previous experiments and known-best configurations to recommend and rank likely good approaches to novel problems.

Both components also consider task runtime, incorporating a general framework of constraints and re-ranking proposed models in a principled manner. This will flexibly incorporate constraints on the search from any source, though we primarily regard the constraints coming from the data, previous experiments, theory, and expert best practices.

3.3. Mixed-Initiative Question-Driven Modeling Workbench

The Distil workflow for building question-driven models guides domain expert users through assembly of data to support answers, nomination of features to consider in modeling answers, and inspection of models and model performance, with means for iterative refinement. To illustrate this workflow, we present a simple scenario in which a user without data science expertise wishes to predict vehicle acceleration using quantitative mileage performance data.

Select Data (Figure 1a). To begin, the user searches Distil's data lake for relevant data by expressing intent in natural language (*predict vehicle acceleration*). Distil returns an annotated list of datasets that contain references to acceleration, for user selection.

Select Target (Figure 1b). Once the *d_196_autoMpg* dataset is selected, Distil lists the features it contains, sorted by the Novelty Detection measure of interestingness. Semantic Data Type Classification infers the type of data in each feature and visualizes the range of values found in the dataset. The user chooses to model and predict *acceleration*.

Create Models (Figure 1c). Next the user begins the model building process by nominating factors to empirically model, from the remaining features. Interactive visualization aids decision making. Highlighting a range in a feature of interest shows its relationship with other features, by highlighting corresponding records. Outliers such as *horsepower=0* can be rapidly identified as bad samples and excluded from modeling. Users with no tacit knowledge may click to add all features without exclusion.

Once features have been selected, the user initiates the creation of models. Distil automatically chooses algorithms to propose several possible models to predict the selected target (*acceleration*) using the features to model (*horsepower* and *cylinders*). Because *acceleration* is scalar, Distil generates regression models.

Review Results (Figure 1d). The View Models stage enables the user to evaluate and compare model results generated by Distil. An Error slider controls the acceptable degree of variance in classifying predictions as correct[1]. As in the previous step, the user can interactively highlight correct or incorrect sample sets within feature summaries to identify contributing factors, and return to a previous step to make refinements.

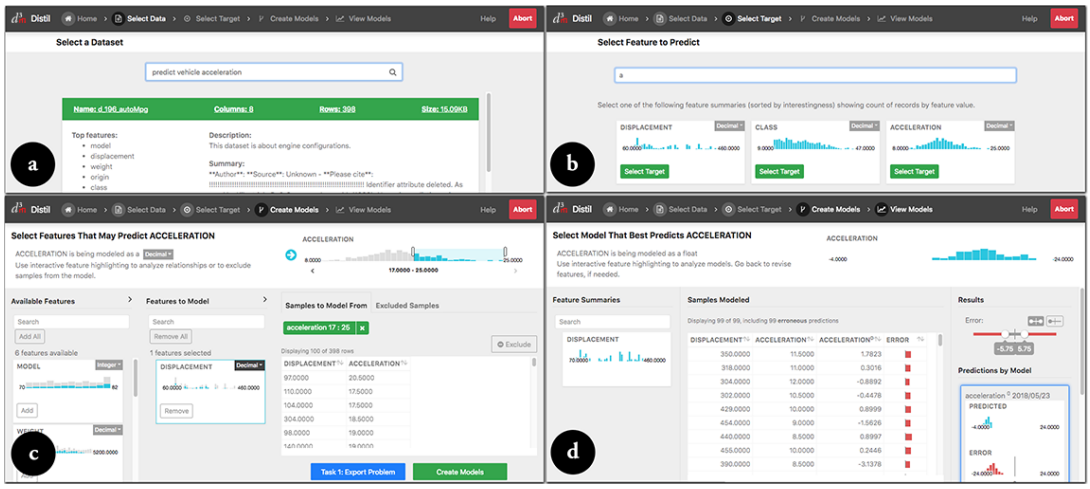

Figure 1: Machine intelligence guides the user through creating a regression model of acceleration, from (a) data selection, to (b) target selection, (c) model configuration and (d) analysis of results.

4. Conclusion

In this paper we presented an overview of Distil, a mixed-initiative system to enable domain experts to conduct question-oriented and data-driven model discovery using an iterative workflow. Internal testing and development was conducted using open datasets from sources such as OpenML (OpenML, 2018) and Kaggle (Kaggle Inc, 2018). Early formal user testing was conducted by NIST (NIST, 2018) using challenge problems on blind datasets have been encouraging. Several areas of future work have been identified and are planned for further experimentation: 1) expansion beyond tabular datasets to include imagery, time series, and graph data; 2) fusion of multiple heterogeneous datasets to expand the domain of potential models; 3) expansion beyond simple models that predict a single value to derive models that can be separated into facets of a larger complex system that are a composition of a hierarchy of models (Jonker, 2012); and 4) interaction and scalable visualization of models for sensemaking, and "what" if analysis (Langevin et al., 2015).

Acknowledgments

This work was supported, in part, by the Defense Advanced Research Projects Agency (DARPA) (contract number D3M (FA8750-17-C-0094)). The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.

1. For categorical features, Distil generates classification models. The accuracy of predictions in these instances is simply whether the predicted value matches the actual value.

References

Mihael Ankerst, Christian Elsen, Martin Ester, and Hans-Peter Kriegel. Visual classification: an interactive approach to decision tree construction. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 392–396. ACM, 1999.

Paul Azunre, Craig Corcoran, David Sullivan, Garrett Honke, Rebecca Ruppel, Sandeep Verma, and Jonathon Morgan. Abstractive tabular dataset summarization via knowledge basesemantic embeddings. *arXiv preprint arXiv:1804.01503*, 2018.

David Ben-David, Tamar Domany, and Abigail Tarem. Enterprise data classification using semantic web technologies. In *International Semantic Web Conference*, pages 66–81. Springer, 2010.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Kristin Cook, Nick Cramer, David Israel, Michael Wolverton, Joe Bruce, Russ Burtner, and Alex Endert. Mixed-initiative visual analytics using task-driven recommendations. In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*, pages 9–16. IEEE, 2015.

Alex Endert, Patrick Fiaux, and Chris North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 473–482. ACM, 2012.

Alex Endert, M Shahriar Hossain, Naren Ramakrishnan, Chris North, Patrick Fiaux, and Christopher Andrews. The human is the loop: new directions for visual analytics. *Journal of intelligent information systems*, 43(3):411–435, 2014.

Christiane Fellbaum, editor. *Wordnet: An electronic lexical database*. MIT Press Cambridge, 1998.

Patrick Fiaux, Maoyuan Sun, Lauren Bradel, Chris North, Naren Ramakrishnan, and Alex Endert. Bixplorer: Visual analytics with biclusters. *Computer*, 46(8):90–94, 2013.

Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 489–500. ACM, 2015.

Philip J Guo, Sean Kandel, Joseph M Hellerstein, and Jeffrey Heer. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 65–74. ACM, 2011.

Jeffrey Heer, Joseph M Hellerstein, and Sean Kandel. Predictive interaction for data transformation. In *CIDR*, 2015.

Heiko Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007.

M Shahriar Hossain, Praveen Kumar Reddy Ojili, Cindy Grimm, Rolf Müller, Layne T Watson, and Naren Ramakrishnan. Scatter/gather clustering: Flexibly incorporating user feedback to steer clustering results. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2829–2838, 2012.

IBM. IBM Watson Analytics, 2018. URL https://www.ibm.com/watson-analytics.

Christopher R Johnson, Charles J Fillmore, Miriam RL Petruck, Collin F Baker, Michael Ellsworth, Josef Ruppenhofer, and Esther J Wood. Framenet: Theory and practice, 2002.

David Jonker. Linked visible behaviors: a system for exploring causal influence. In *2nd International Conference on Cross-Cultural Decision Making*. AHFE, 2012.

Kaggle Inc. Kaggle datasets, 2018. URL https://www.kaggle.com/datasets.

Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2917–2926, 2012.

Scott Langevin, David Jonker, Kevin Birk, Chris Bethune, and Nathan Kronenfeld. Global to local pattern of life analysis with tile-based visual analytics. In *IEEE Vis 2015, Practitioner Session*, 2015.

Stephen Makonin, Daniel McVeigh, Wolfgang Stuerzlinger, Khoa Tran, and Fred Popowich. Mixed-initiative for big data: the intersection of human+ visual analytics+ prediction. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on*, pages 1427–1436. IEEE, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

New Knowledge. PUNK - Primitives for Uncovering New Knowledge, 2017. URL https://github.com/NewKnowledge/punk.

New Knowledge. NewKnowledge/simon: Character-level CNN+LSTM model SIMON - Semantic Inference for the Modeling of ONtologies - for text classification, 2018. URL https://github.com/NewKnowledge/SIMON.

NIST. National Institute of Standards and Technology, 2018. URL https://www.nist.gov/.

OpenML. OpenML, 2018. URL https://www.openml.org/.

OpenRefine. OpenRefine GitHub, 2018. URL https://github.com/OpenRefine/OpenRefine.

Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *International Symposium on Smart Graphics*, pages 184–195. Springer, 2010.

Pedro Szekely, Craig A Knoblock, Shubham Gupta, Mohsen Taheriyan, and Bo Wu. Exploiting semantics of web services for geospatial data fusion. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatial Semantics and Ontologies*, pages 32–39. ACM, 2011.

Xu-Meng Wang, Tian-Ye Zhang, Yu-Xin Ma, Jing Xia, and Wei Chen. A survey of visual analytic pipelines. *Journal of Computer Science and Technology*, 31(4):787–804, 2016.

Jia Yue, Anita Raja, and William Ribarsky. Predictive analytics using a blackboard-based reasoning agent. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 2, pages 97–100. IEEE, 2010.