Automating Data Exploration through "Interestingness" and Insights

Jing Shyr, Damir Spisic
IBM SPSS Predictive Analytics, 200 W. Madison St. 23th Fl., Chicago, IL 60606

Abstract

Data exploration includes generating an overview of variables contained in a dataset and an understanding of the most important relationships among the existing variables. Data needs to be prepared for modeling while the models themselves are used for gaining insights into the main aspects of the discovered relationships. It involves the use of many statistical techniques that require the judicious application of a proficient analyst. The task becomes much more difficult for large datasets with many variables. We present a robust automation framework that runs a set of exploratory statistical analyses on a given dataset. The statistics range from univariate analyses establishing the necessary metadata information to multivariate analyses discovering the relationships between the target variables and potential predictors. All the results are sorted by a new "interestingness" index comprised of suitable statistics. This allows the most relevant relationships to be selected. The most important aspects of the discovered relationships are further analyzed and presented as insights and expressed using a non-technical language allowing non-experts to gain a better understanding of their data.

Key Words: Data exploration, automation, interestingness, insights

1. Introduction

This work was motivated by a combination of several emerging trends relevant for the exploratory data analysis of big data.

Thanks to the high volume production of electronic devices capable of generating, distributing and collecting data and their quick penetration into all areas of work and life, new data is being generated daily at a high rate. The amount of stored data increases beyond the rate of increase in computing capacity. "Big data" is the term that has been introduced to describe the nature of this emerging data. It is typically very large in size and may be stored in multiple files and locations or even in streams due to a lack of capacity available for being stored. It can be structured or unstructured using many different data formats and metadata descriptions. When represented by a tabular structure, it contains a very large number of records, or very large number of variables or both.

Powerful and flexible computer clusters are becoming ubiquitous as well. They allow for a scalable increase of processing power by simply adding machines to the cluster. Their capacity is often measured in the amount of electric power used rather than in the speed of their processors. The key scalability issue is the capability of adding machines to the

system in order to be able to process increasing amounts of data in approximately the same amount of time.

This scalability requirement on computing resources places a related requirement on the algorithms used for data analyses. Many of the data analysis algorithms require processing times that grow faster than the amount of data to be analyzed. This makes them impractical for very large data sets as they do not allow for solving larger problems by simply adding more computational power. More efficient algorithms are needed that are able to exploit the advantages rendered by the increasingly available computer clusters.

Last but not least, there is a human factor involved as well. Propagation of data analyses into larger numbers of organizations and across different departments within these organizations requires additional skilled analysts to make sense of this information and present the results to decision makers. Based on historic division of labour this generates unrealistic demands on the number of analysts needed. A different approach is needed that enables non-experts to execute data analyses and draw relevant conclusions. Moreover, exploring data with large number of variables can be taxing even for proficient data scientists.

Our proposal addresses computational efficiency for exploratory analysis by executing large number of smaller analyses in parallel. When necessary, new algorithms are developed that ensure scalability appropriate for the volume of data. Automated analysis makes an efficient use of computational resources while resulting in a large number of smaller sets of results. Ranking the generated results and showing the ones that are more relevant allows for effective and productive exploratory analysis of data. Automatically generated insights are lowering the skill threshold and allowing larger audience to get directly involved in the data analysis.

In Section 2 we present the Automation Analysis approach to data processing. In the Section 3 we discuss our overall approach to Automated Data Exploration. Section 4 describes the Metadata Discovery process. Section 5 and Section 6 describe the interestingness indices and generation of insights respectively. In Section 7 we offer conclusions and future directions.

2. Automated Analysis

Big data present specific challenges when it comes to exploratory data analysis. Distribution summary and computation of other summary statistics for a single variable can be executed very efficiently for most statistics. Similar is true for any given bivariate association as well as relationships between a target and some number of given predictors. However, performing a single analysis still requires a data pass over very large and possibly distributed data. The established approach of judiciously selecting analyses to perform and explore data in the sequential manner that depends on previously obtained results quickly becomes infeasible.

Automation brings efficiency to this process. All variables are summarized, appropriate transformations are determined and all bivariate analyses are executed. Afterwards, numerous model-based analyses of predictor influence on the given target variables are generated. This occurs with minimal or no user intervention. While discovery of metadata

properties is largely automatic, users can provide additional information that provides necessary focus and relevancy to the performed analysis.

The key to efficient processing is the parallel execution of a large number of analyses. This reduces the number of costly passes over very large data sets and takes the advantage of the scalable cluster architecture. Figure 1 shows the data flow for the computation steps supporting this approach. Descriptive Engine computes all univariate statistics in the first data pass and bivariate statistics in the second step. A subsequent Interaction Detection module generates Key Drivers in two data passes. Finally, a Decision Tree module requires as many data passes as there are levels in the model. Creating only a few levels of the tree model is usually sufficient for exploratory purposes. Other suitable models can be easily added within this framework.

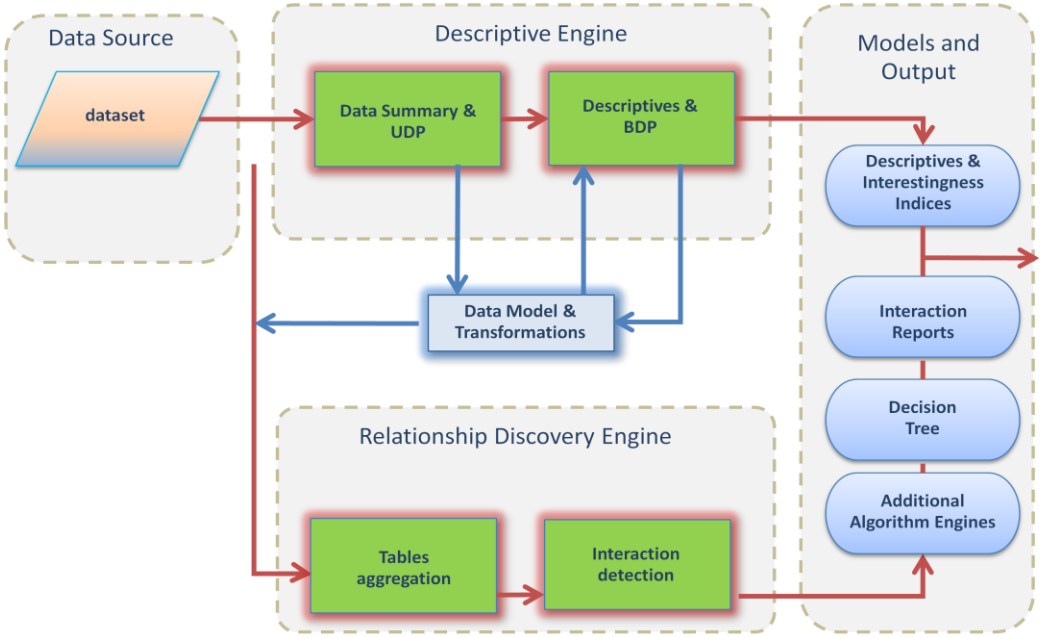

Figure 1: Architecture and data flow for computational modules supporting automation.

In addition to executing many analyses, parallelism is also exploited within the algorithms for efficient processing of large and/or distributed data. Algorithms operate on data blocks and generate intermediate results that are subsequently merged and the analyses for the overall data set are computed. For most univariate and bivariate statistics, development of parallel algorithms is quite straightforward. An important exception is order statistics which require a different approach. Storage of data in memory and sorting the data values is no longer a viable option for very large data sets. Likewise, taking advantage of parallel processing and data distribution is a challenge when it comes to creating a large number of small models in the Interaction Detection module and overall models such as decision trees.

There are several relevant results in designing parallel algorithms that can be used for the automation purposes and run on computer clusters capable processing very large data sets. For example, some are related to computing order statistics as described by Chu et al. in [2], while others are related to imputing missing data given by Shyr et al. in [3]. A scalable approach to interaction detection for data with large number of predictors is

discussed by Shyr et al. in [4]. Other parallel algorithms for common statistical methods and models have been developed as a result of vigorous research in this area in recent years.

3. Automated Data Exploration

Upon obtaining a large number of analysis results appropriate for the given data, the challenge is to present them to the user in a coherent and understandable display. The most relevant analyses need to be easy to interpret. This obvious rule has led us to develop a measure which we call the "interestingness" index.

An interestingness index is a summary measure for a given analyses. Its values are between 0 and 1 where 0 stands for the least interesting and 1 for the most interesting analysis. We compute an interestingness index for each analysis type so that all similar analyses can be compared and ranked. For example, in the first release of IBM SPSS Analytic Catalyst we consider four types of analyses: Univariate summaries, Bivariate summaries, Key Drivers and Tree models. Using corresponding interestingness indexes, all the variables in the data set are ranked according to their properties of the corresponding Univariate summaries. All the pairs of variables are ranked and compared using an interestingness index for Bivariate summaries. A separate interestingness index was developed for ranking the models generated in the Interaction Detection module that supports Key Drivers detection. Specifying multiple target fields leads to the creation of multiple Tree models. They are also ranked according to an interestingness index designed for Tree models specifically.

Another element of automated exploration process is the notion of "insight". Insights are derived for each conducted analyses and presented to the user in plain language. Insights interpret the statistical results to users lacking the necessary knowledge and experience to draw conclusions on their own. This is achieved by using thresholds and dividing range of an interestingness into intervals where each interval of values receives a common interpretation. Insights help analysts to evaluate different aspects of each available analysis.

The final exploration design element is guiding users through the analysis results. Guidance achieved through visual and general output design means. Detailed properties for each analysis are available through the links describing the contents so that the user can easily "drill down" to the properties of interest. Dynamic Visual Insights are available for charts when appropriate in order to progressively display particular analysis properties in the context of the analysis as a whole. Together with statistical results and worded insights this completes the automated data exploration support making the data exploration process much simpler for a broad audience.

4. Metadata Discovery and Data Transformations

Efficient algorithm parallelization and execution of many concurrent analyses is an important aspect of automating data exploration. Most of the applicable analyses require some basic information about data that can be described as metadata properties. Metadata discovery is the automated process designed to obtain this information.

A critical step in metadata discovery is determining metadata properties for each variable in the data. After the first data pass we obtain Univariate summaries for all fields that depend on variable being either numeric or string variable. We use all the available variable summaries to determine variable measurement level and role in addition to its utility in further analyses.

This information is generated automatically, but the user input is sometimes necessary and can also provide additional focus for the analysis. Figure 2 shows the process where automated discovery results are inspected by the user. Metadata properties are adjusted and summaries are refreshed until the user is satisfied with the specifications.

Specifying metadata properties can be viewed as a chance for the user to provide their input into exploratory analysis. This sometimes requires a degree of understanding of the nature and domain of the data as well as the analysis goals. For example, user may specify a target variable. However, no insight into the subsequent analyses and modeling is required. Indeed, while the user may not be a data analyst, he or she would typically have some understanding of the source and nature of the data that they are interested in analyzing.

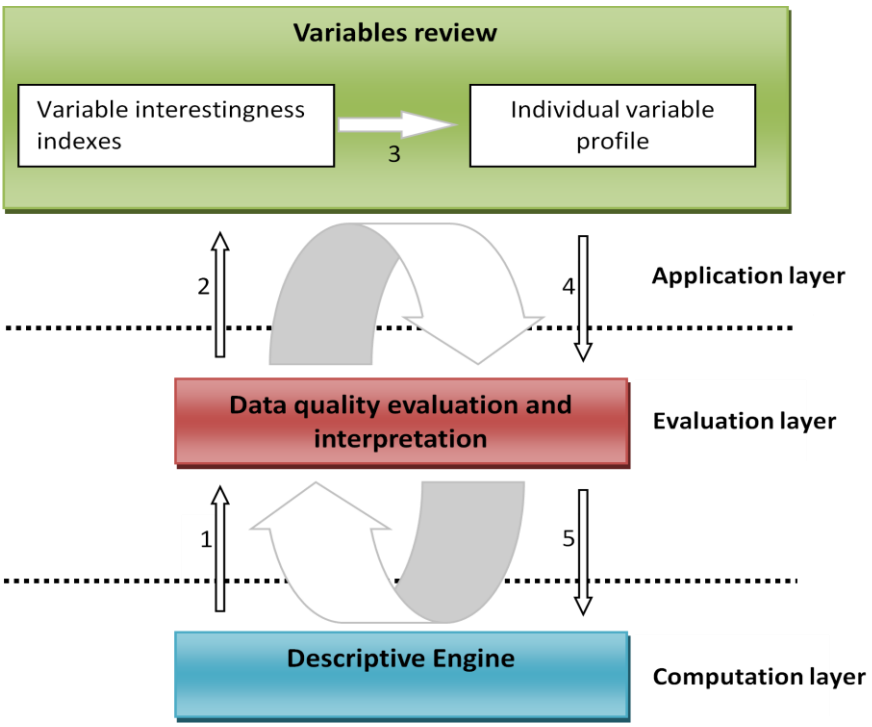

Figure 2: Process for specifying metadata: automated discovery and user interaction.

Metadata properties include variable role, variable measurement level, missing data values specification and outlier handling specification. In addition, various variable transformations are generated that may also need to be accepted by the user. Examples of transformations are imputation of missing data, supervised merging of categories, handling of outliers and transformation to symmetry. These are all useful data preparation tools, but user input may be needed to approve their inclusion in subsequent analyses.

5. Interestingness

Upon obtaining possible user input relevant to the generated metadata properties, the remaining data exploration runs without any additional user intervention. Simply, all relevant exploratory analyses are executed based solely on the metadata specifications. The challenge is for the user to obtain and view the analysis results that are relevant and insightful for a given data set.

Interestingness is introduced precisely for this reason: to rank the results within each analysis type so that important analyses of interest to the user can be easily reached. Users may also be interested in some particular analyses, but having the overall view of which analyses are more relevant makes the exploration a lot easier, reliable and indeed possible when the number of available analyses becomes very large.

Interestingness is a standardized measure of analysis relevance for each analysis type used for data exploration or modeling. It is based on one or several sub-indices where each sub-index is based on a statistic or a set of statistics that contain relevant information about given analysis. The index and composite sub-indices are always standardized so that their values are between 0 and 1. This makes it possible to compare and contrast all analyses within the same analysis type even when the analyses methods are different.

Overall index	**Main index**	**Sub-index**
Overall interestingness for **categorical** variables	Special values	Completeness
		Influential category
	Shape	Imbalance
Overall interestingness for **continuous** variables	Special values	Completeness
		Outliers
	Shape	Asymmetry
		Peakedness
		Distribution shape

Table 1: Interestingness index and sub-indices for univariate analysis.

For example, Table 1 shows the interestingness indices and corresponding sub-indices for univariate analyses of variables having either continuous or categorical measurement level. Overall interestingness index for categorical variables is specified using three sub-indices, while the overall index for continuous variables is specified using five sub-indices. Each sub-index is based on a single univariate statistic. Sub-indices can also be grouped into main indices as shown.

Sub-indices are specified by a given statistic and a monotone transformation mapping the range of the corresponding statistic onto the unit interval. An example of the transformation that can be used is a smooth piecewise polynomial for statistics with bounded range and coupled with an inverse exponential for the unbounded part of the

range. Such transformations make the statistics values standardized so that they can be easily combined and compared across analyses.

This also makes it straightforward to combine a number of sub-indices into an overall interestingness index for the given analysis. A combination function that combines the values of sub-indices into the overall index can be average, maximum or weighted average allows for a flexible design with room for customization. Some aspects of the given analysis may appear more relevant in some situations than others. By specifying the corresponding combination weights, the user can place more emphasis on the properties that are more relevant to their purpose.

6. Insights

Users may select a particular set of analysis results based on a high value of interestingness index, or directly identify the analysis by selecting the variables involved. In either scenario, corresponding analysis insights are always generated in plain language describing the analysis results obtained. Generated insights provide a quick overview of the analysis of interest and help the user to understand and interpret the obtained results.

Insights are tied to sub-indices computed for the given analysis. They describe the degree of relevance for each underlying statistical result. The range of underlying statistic for a given sub-index is divided into three sub-ranges of values based on their impact. The ranges can be described as corresponding to "low", "medium" or "high" impact. The threshold values between the ranges are selected based on the experience and represent standard statistic thresholds for interpreting values of the particular statistic. When statistic values are mapped into the corresponding standardized sub-index values, the statistic thresholds are mapped into the corresponding sub-index thresholds. The sub-index thresholds are normally selected to be the same across all sub-indexes for a given analysis. For example, 1/3 and 2/3 can be commonly used. This further facilitates and standardizes reporting of the results. If the sub-index values are in the specified standard sub-range, appropriate insight can be issued without any additional information. Notice that the insights are not based on the combined overall interestingness indices. They lead to understanding of the components that combine into the overall index.

6.1 Example

In this example we demonstrate computation of a sub-index and setting the thresholds suitable for a bivariate association strength analysis. We assume that variables are categorical and use Cramer's V as the underlying statistics. Set 0.2. and 0.6 as threshold values. The interpretation of resulting intervals for Cramer's V is given in Table 2.

Cramer's V	**Interpretation**
< 0.2	weak
0.2 – 0.6	moderate
>0.6	strong

Table 2: Example of thresholds for interpreting Cramer's V statistic.

The statistic is mapped into a sub-index by using cubic polynomials smoothly tied at the threshold mapping points. Cramer's V threshold values 0.2 and 0.6 are mapped into the sub-index thresholds 1/3 an 2/3.

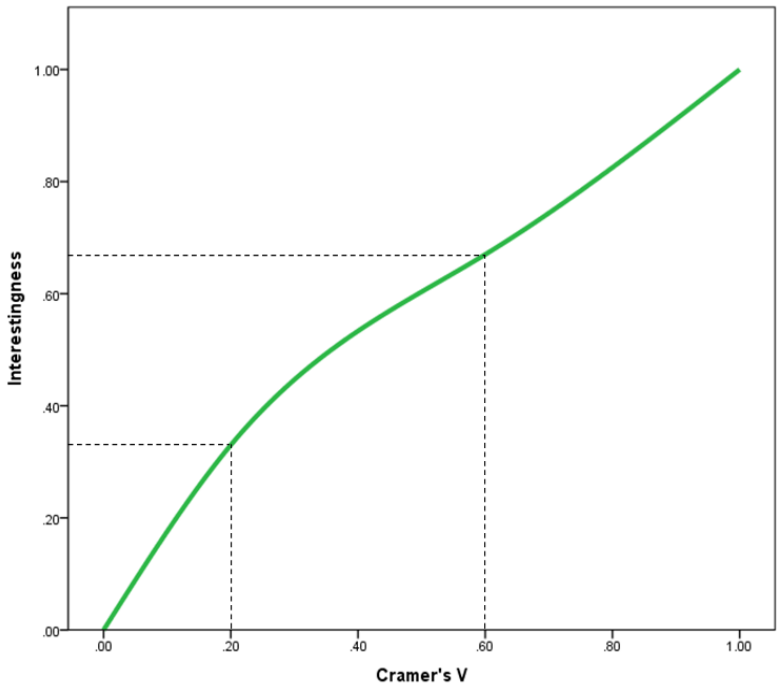

Figure 3: Mapping of Cramer's V statistics into interestingness index with specified thresholds.

Assuming that the value of sub-index for variables "Education" and "Position level" is 0.7 for the given data, we can report to the user that "Association strength between Education and Position Level is high."

Additional details and examples on specifying interestingness index for univariate analyses can be found in document [1] by Shyr et al.

7. Conclusions and Future Work

Data exploration for big data is an important and challenging problem. Our solution proposes the use of automation to address computational complexity of the overall process. This approach facilitates the analysis and provides for interactive and guided selection of results by users. It allows them to quickly spot relevant fields, bivariate associations and key drivers for the specified target. The obtained results are explained and easily comprehended by the broad audiences while making it easier for experts to sift through the large amounts of information.

We will continue to explore this approach on a wider set of exploratory analyses, transformations and different modeling techniques appropriate for different types of data. While this approach may not provide detailed and complete analysis for each and every problem, it clearly provides a quick and easy way of exploring critical data properties,

providing some insights and thereby guidance to more specific problems to be further addressed by more focused techniques.

References

[1] Shyr, J, Spisic, D., Wright, R., Xu, J., and Zhang, X.Y. "Interestingness of Data", US 2013/0007003 A1.

[2] Chu, J., Han, S., Li, F., Shyr, J., Spisic, D., Wills, G. and Xu, J. "Computing and applying order statistics for data preparation", US 2013/0218909 A1.

[3] Shyr, J., Chu, J. and Han, S. 2012. "Missing value imputation for predictive models on large and distributed data sources". In *JSM Proceedings*, Social Statistics Section. Alexandria, VA: American Statistical Association. 5134-5145.

[4] Shyr, J., Spisic, D., Chu, J., Han, S. and Zhang, X.Y. 2012. "Model-based and scalable relationship discovery in business analytics". In *JSM Proceedings*, Social Statistics Section. Alexandria, VA: American Statistical Association. 5146-5158.