International Workshop on Visual Analytics (2011)
S. Miksch and G. Santucci (Editors)

SmartStripes - Looking under the Hood of Feature Subset Selection Methods

T. May[1], J. Davey[1] and T. Ruppert[1]

[1]Fraunhofer Institute for Computer Graphics Research, Germany

Abstract

We propose a visualization method for the diagnosis and interactive refinement of automatic techniques for feature subset selection. So-called filter techniques use statistical ranking measures to identify the most useful combination of features for further analysis. Usually a measure is applied to all entities of a data-table. The influence of atypical entities can distort the result, but this distortion may be masked by the statistical aggregation. Clearly, feature and entity subset selection are highly interdependent. Our technique, SmartStripes, intends to make this interdependency visible.

1. Introduction

Ana wants to use a clustering algorithm to cluster a high-dimensional data set for her boss. The data is delivered to her as a data table with a high number of *features* (columns) and *entities* (rows). Ana has heard of the so-called curse of dimensionality; she knows that her clustering algorithm probably won't deliver useful results if she uses all of the available features as input. Ana usually uses a representative subset of the available features for clustering.

To aid her in her search for a suitable feature subset, Ana runs a *feature subset selection* algorithm. This algorithm attempts to select the smallest possible subset of features with the highest possible correlation with the non-selected features. Ideally, this subset should contain the most relevant features for the clustering task. Additionally, the selected features should exhibit almost no dependencies on one another; because dependency implies redundancy and redundancy in the input can skew the results of her clustering algorithm.

Ana runs the feature subset selection algorithm, then feeds the selected subset into the clustering algorithm and sends the results to her boss. A couple of days later, her boss sends her an email asking why certain apparently important features were not included as input for her clustering algorithm. Ana retraces her steps and takes a closer look at the data set to find the reason why the features were not selected. She discovers that one third of the entities in the data set contain the value *no data* in important features. These values led to an artificial increase in the dependencies between these features. Which, in turn, quite rightly led to their exclusion from the selected feature subset.

This example illustrates a rather trivial case. In general the results of feature subset selection algorithms are dependent on the entity subsets on which they operate. By focusing on the other two thirds of the data (i.e. ignoring the *no data* entries), Ana would be able to correct the error and achieve more useful clustering results. However, there are frequently less obvious dependencies, which lead to equally problematic skewing of clustering results.

One naive solution would be to run feature subset selection algorithms simultaneously on a number of different entity subsets. The results of each run could then be compared to find the *best* feature subset. Due to the fact that entities usually far outnumber features, this approach would simply cause an explosion in computational complexity.

In this paper we propose the use of a suite of visualization techniques called *SmartStripes* to support the diagnosis of problems with and the interactive refinement of feature subset selection. In particular, we show how *SmartStripes*

DOI: 10.2312/PE/EuroVAST/EuroVA11/013-016

allows analysts like Ana to explore complex dependencies at the entity subset level.

2. Related Work

The selection or generation of a feature subset from a high-dimensional table is considered a canonical step in the knowledge discovery process [FPSS96]. Given N features, there are $2^N - 1$ possible feature subsets to choose from. Thus Guo [Guo03] argues that automatic methods are an indispensable aid to feature subset selection. Guyon and Elisseeff [GE03] present a survey including an extensive description of the caveats and pitfalls of automated feature selection. They outline the general problem as finding a *minimal subset* of features, which together are *most useful* for the following analysis steps. They claim that no single method can be expected to find the best feature subset in all cases.

Kriegel et. al [KKZ09] propose a terminology for the categorization of approaches for the clustering of high-dimensional data. They identify *local feature relevance*, i.e. the manifestation of different clusters in different feature subsets, as one of the problems of this field.

Feature subset selection methods can be divided into three categories; filters, wrappers and embedded methods [GE03]. Filter methods use statistical ranking criteria for the evaluation and selection of feature subsets. Wrapper methods use quality measures of clustering (or other data-mining) techniques [KJ97], while in embedded methods the feature selection is intertwined with clustering or other data-mining algorithms. In this paper we will focus on the support of filter methods.

Guo [Guo03] argues that human intervention is necessary to evaluate and guide the procedure of feature subset selection. Most visual support of the task of feature selection is either on a very coarse level of detail (e.g. a correlation matrix) or on a very fine level of detail (e.g. scatterplot matrices or parallel coordinates). Highly detailed views are not suitable for feature subset selection because they do not scale well with the number of features. To our knowledge, the visualization most commonly used is a correlation matrix. Correlation matrices show statistics for all binary correlations between the features of a given table. Thus, correlation matrices condense the relationships between feature pairs to single values. Friendly [Fri02] presents an in-depth exploration of the design space of correlation matrices.

MacEachren et al. [MDH*03] and Ingram et al. [IMI*10] each present a framework which includes an interactive feature selection step in its respective analytical process. Both use a correlation matrix to display bivariate dependencies. While they do provide a space efficient overview, we feel that the correlation matrices limit the analyst's options for problem diagnosis and interactive refinement. Ingram et al. use Pearson's correlation coefficient as a measure of correlation between feature pairs. MacEachren et al. use maximum conditional entropy. Like other measures that apply to the distribution of values rather than the values themselves, entropy measures have the advantage of being applicable to nominal, ordinal and discretized numerical features.

Yang el al. [YWRH03] present a feature selection method which is based on agglomerative hierarchical clustering. The process can be interactively controlled in order to identify a meaningful and useful feature subset. Their approach, however, relies on the definition of similarity measures for all pairs of features. With nominal or ordinal attributes in the data table this requirement is almost never fulfilled.

Johansson et al. [JJ09] present an approach to tackle the *competing* structures in a data set, which are emphasized or masked depending on the quality metrics chosen. They propose a user-defined mixture of different statistics to control the dimension reduction process. We present an orthogonal approach, because we are focussing on the interdependency between entity and feature subset selection.

3. Approach

We define the following requirements for a visualization to support feature subset selection:

1. Provide an overview over as many features as possible; ideally the visualization should be scalable with the number of features.
2. Show the details of dependence between whole features and between entity subsets of those features.
3. Use the same measures for all feature types; do not use separate measures for nominal, ordinal and numerical data.
4. Support the diagnosis of problems and the interactive refinement of the results of automated techniques.

To meet these requirements, we developed the *SmartStripes* suite. At present, the suite consists of two components; the Feature Partition View and the Dependency View. In the following subsections we will describe these two views in detail.

3.1. Feature Partition View

With the Feature Partition View (see Figure 1 (left)) a discretization of the features, needed for the generation of the Dependency View (see section 3.2), the main visualization component of *SmartStripes*, is defined. Each feature is visualized as a histogram, where its values are grouped into buckets or clusters. The initial clustering or bucketing of the numerical, ordinal or nominal features is automatically computed, as described in the following.

We chose the k-medoids clustering algorithm for the discretization of our numerical features. The choice of k-medoids is justified by its simplicity and by its tolerance of outliers. The number of clusters, k, is chosen to guarantee a

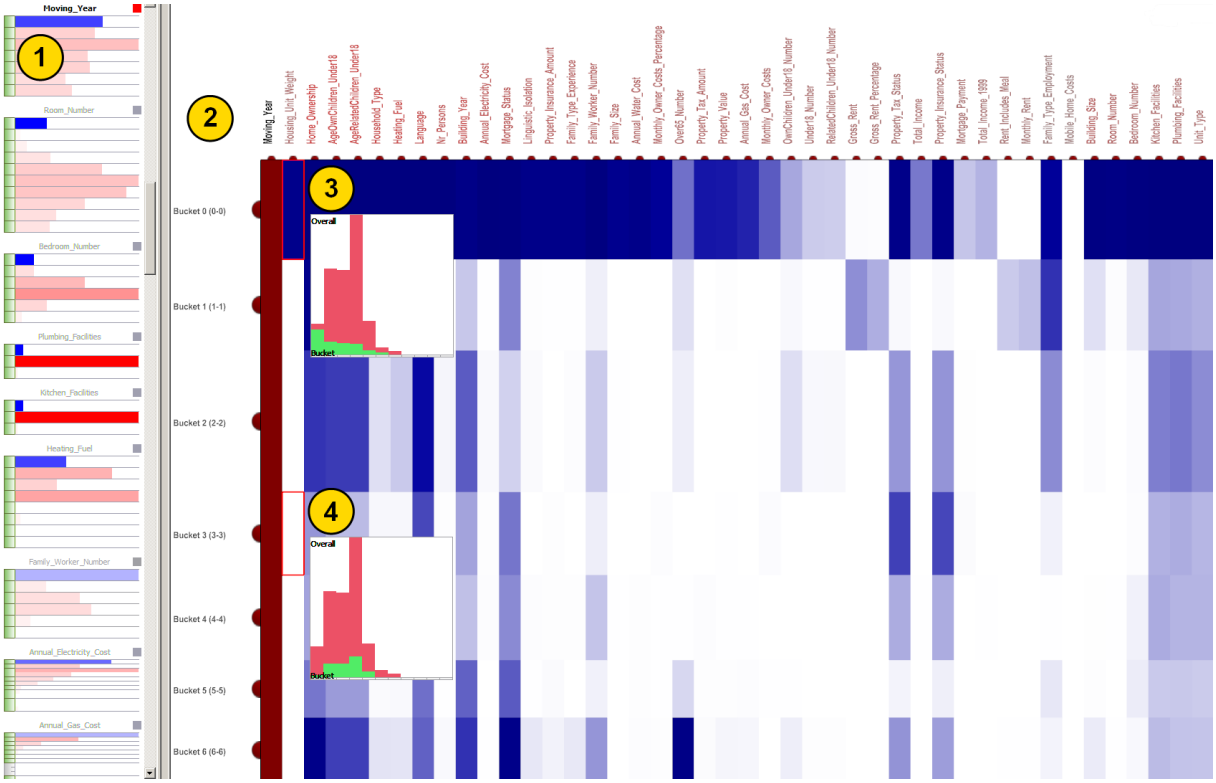

Figure 1: *Part of the US-microcensus data set shown with SmartStripes. Feature Partition View (1.): Histograms visualizing the feature discretizations. The user can interactively reorder, regroup or delete buckets. Dependency View (2.): Columns represent features, rows represent entity subsets defined by a selected feature (red column). The height of the rows is proportional to the number of entities in the corresponding entity subsets. Note that the statistical measures along a column are most often not homogeneous. Highly saturated cells (3.) indicate a high dependency on the selected feature for the corresponding subset. The detailed view contrasts the overall distribution (red histogram) with the distribution in this bucket (green histogram) of feature values. The greater the difference between the distributions the higher the dependency. The opposite is true for less saturated cells (4.). The blue row on top indicates a strong influence of the first bucket over virtually all features, which might distort the result of automatic techniques.*

minimum average number of entities in each cluster. This is an important prerequisite for reliable results of our chosen measure.

Ordinal and nominal features are bucketed into value ranges if the overall feature ranges are too large. Here again, a minimum number of entities per bucket determines what we mean by *too large*.

After the initial discretization or bucketing step, the user is able to interactively refine this clustering or bucketing in the visualization. Possibilities for refinement include redefining value ranges for ordinal and numerical attributes, regrouping of nominal attribute values and the deletion of whole buckets. This is a very important opportunity for the user to incorporate her knowledge into the analysis process and goes some way to satisfying the fourth design requirement specified at the beginning of this section.

3.2. Dependency View

The main visualization component of Smartstripes is the Dependency View. It shows the dependencies between a selected feature and all other features in the data set. A heat map display is used, in which the features are shown as columns and entity subsets are shown as rows (see Figure 1 (right)).

The rows of the heat map represent the buckets or clusters of the selected feature, defined in the Feature Partition View (see section 3.1). In the following description, we will restrict our discussion to buckets, the same holds, of course, for clusters. The saturation of each cell is determined by the statistical measure. We chose a variant of Pearson's Chi-Square test (also known as the Chi-Square Goodness-of-Fit test) for the determination of saturation. This measure could be replaced by another appropriate test. The test produces

a statistic for the comparison of two distributions. The distributions used for the test are estimated by the frequency counts of values in each bucket considered, normalised using the total number of values. For each feature and each bucket determined by the discretization of the selected feature the distribution in the bucket is compared with the overall distribution of the feature. If the distributions are almost the same, then the correlation with the bucket of the selected feature is weak; this leads to a low saturation. If the difference is large, then a dependency is probable, leading to a high saturation.

The formula for our Chi-Square test measure is given below.

$$\chi^2 = \sum_{i=1}^{n_1} \frac{(O_{i1} - E_{i1})^2}{E_{i1}} + \cdots + \sum_{i=1}^{n_m} \frac{(O_{im} - E_{im})^2}{E_{im}} \quad (1)$$

Each summand determines the saturation value for its corresponding cluster or bucket represented by a cell in the visualization (see figure 1). Adding all the summands up would result in a single statistic for the correlation of two features. The O_{ij} represent the distribution of the test feature values in the bucket or cluster j of the selected feature. The E_{ij} represent the distribution of the values in the full test feature range.

By visualizing not the single statistic, but its components, we achieve a somewhat finer level of detail. The chosen visualization technique (i.e. the heat map) allows the simultaneous visualization of a large number features. Thus, the Dependency View fulfils the first two requirements specified at the beginning of this section. In addition, the use of a measure applicable to all data types satisfies the third design requirement.

The user interacts with SmartStripes by selecting a feature (i.e. a column). The dependency measures between the selected feature and all other features will be displayed. In addition the user may exclude or include entity subsets by toggling the corresponding rows. The statistical measures will be applied to the active part of the data only.

In addition to manual selection, an automatic filter method can be used. Instead of computing the complete subset in one single run, our implementation adds or removes features step by step. Selected features are highlighted for further inspection or manual modification. Currently, we use a feature ranking scheme based upon the Chi-Square measure.

4. Conclusion & Future Work

Ana is able to inspect the details of the measures used for the feature selection. She steps into the process and checks the dependencies between a new feature and the remainder of the data table. Ana notices a relatively strong dependency which encompasses almost all features, but only a subset of the data (see figure). The documentation reveals that this is an obvious relationship - at least for the human user. Hence, she feels safe in removing this subset from her analysis to get a *cleaner* data set for her clustering algorithm.

We presented *SmartStripes*, a visualization technique for the in-depth inspection of feature subset selection methods. Instead of aggregating the dependency between two attributes to a single measure, we decompose the statistical aggregation by entity subsets and show their individual contribution to the measures. Our technique is work in progress: among the issues that are still to be adressed are feature sorting. This is a major concern in [Guo03], [IMI*10] and [JJ09] and we expect it to be helpful if the number of features exceeds the screen space. Another important issue is *high-level* guidance for the user; in order to ease the interpretation of the visualization depending on her task. Finally, *SmartStripes* will be embedded in a framework and used to steer other techniques for clustering and classification.

References

[FPSS96] FAYYAD U. M., PIATETSKY-SHAPIRO G., SMYTH P.: From data mining to knowledge discovery in databases. *AI Magazine 17*, 3 (1996), 37–54. 2

[Fri02] FRIENDLY M.: Corrgrams: Exploratory displays for correlation matrices. *The American Statistician 56* (November 2002), 316–324. 2

[GE03] GUYON I., ELISSEEFF A.: An introduction to variable and feature selection. *Journal of Machine Learning Resarch 3* (2003), 1157–1182. 2

[Guo03] GUO D.: Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization 2* (2003), 232–246. 2, 4

[IMI*10] INGRAM S., MUNZNER T., IRVINE V., TORY M., BERGNER S., MÖLLER T.: Dimstiller: Workflows for dimensional analysis and reduction. In *Proceedings of the 5th IEEE Conference on Visual Analytics in Science and Technology (VAST)* (October 2010), IEEE Computer Society. 2, 4

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics 15* (November 2009), 993–1000. 2, 4

[KJ97] KOHAVI R., JOHN G. H.: Wrappers for feature subset selection. *Artificial Intelligence 97*, 1 (1997), 273–324. 2

[KKZ09] KRIEGEL H.-P., KRÖGER P., ZIMEK A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data 3* (March 2009), 1:1–1:58. 2

[MDH*03] MACEACHREN A. M., DAI X., HARDISTY F., GUO D., LENGERICH E.: Exploring high-d spaces with multiform matrices and small multiples. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)* (2003), IEEE Computer Society. 2

[YWRH03] YANG J., WARD M. O., RUNDENSTEINER E. A., HUANG S.: Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the symposium on Data visualisation (VisSym)* (Aire-la-Ville, Switzerland, Switzerland, 2003), Eurographics Association, pp. 19–28. 2