Determining Data Relevance using Semantic Types and Graphical Interpretation Cues

Eduardo Haruo Kamioka, André Freitas, Frederico Caroli, and Siegfried Handschuh

{Haruo-Kamioka, Andre.Freitas, Frederico.Caroli, Siegfried.Handschuh}@Uni-Passau.de
Universität Passau

Abstract. The increasing volume of data generated and the shortage of professionals trained to extract value from it, raises a question of how to automate data analysis processes. This work investigates how to increase the automation in the data interpretation process by proposing a relevance classification heuristic model, which can be used to express which views over the data are potentially meaningful and relevant. The relevance classification model uses the combination of *semantic types* derived from the data attributes and *visual human interpretation cues* as input features. The evaluation shows the impact of these features in improving the prediction of data relevance, where the best classification model achieves a F1 score of 0.906.

Introduction

The growing availability of data brings the demand for methods to support the automation of the data interpretation process, by automatically exploring the search spaces of possible interpretations associated with the available data. However, methods to support the automation of large-scale exploratory data analysis are still limited.

The materialisation of the vision of an *automated data analyst* requires a *heuristic model* which can optimise the exploration of the potential interpretation space of the data, detecting which data views and patterns are meaningful and potentially relevant for data consumers.

This work aims at addressing this problem by proposing a *relevance classification* approach based on the composition of *semantic types* and *visual data interpretation cues*. The main goal of the model is to provide a heuristic model which can be used for pruning the search space associated with the interpretation and identification of patterns of interest in the data.

The heuristic model is built upon the assignment of *semantic types* to data attributes which, in combination with *visual interpretation cues*, define a data relevance classification model. Both semantic types and visual interpretation cues are input as features in order to build the final *data interpretation relevance classifier*.

The proposed model lies on the intuition that the semantic types associated with attributes can be used to infer their compatibility to form a meaningful *data view*. Additionally, coarse-grained visual interpretation cues over the final visualisation output (mediated by a specific visualisation type) are used as evidence to detect salient potential patterns of interest within the data. We assess the human interpretation process by systematically and manually classifying meaningful and relevant data views for different domains.

The contributions of this work are: *(i)* the definition of a data interpretation relevance model based on the combination of semantic types and visual interpretation cues; *(ii)* an evaluation of the proposed model and of the impact of semantic and visual cues and *(iii)* the determination of the best classification model through a systematic analysis of different classifiers.

1 Related Work

This work concentrates on the area of automated and intelligent data analysis. In Grosse *et al.* (2012); Duvenaud *et al.* (2013); Lloyd *et al.* (2014) the concept of an *automatic statistician* is introduced. The automatic statistician framework introduces a process to explore the compositionality of a large space of models structures to find the applicable model to predict, classify or extrapolate based on new unseen data. Our approach differs as we explore the compositionality of data views and visual patterns to classify data relevance. Another proposed model is AIDE, which provides a semi-automated process, which relies on planning data analyses steps by a determined combination of data type and user interaction (St. Amant et Cohen (1998); St Amant et Cohen (1997)). AIDE limits its application as a fully autonomous system, requiring corrections executed by the user without training the system to correct itself automatically. Our approach focuses on an automatic classification approach for the selection of relevant data views.

Regarding exploratory data analysis, two works are considered. The first work focuses on automated knowledge discovery workflow composition through ontology-based planning (Záková *et al.* (2011)). It differs from our approach in the semantic representation model where the extraction of semantic features from WordNet hypernyms and distributional word vectors target a more generic semantic representation solution (open vocabulary). The proposed model in this work builds upon Bremm *et al.* (2011) which focused on assisted data descriptors selector based on visual comparative data analysis. It aims at facilitating the user's access to the data analysis process. This data is used to link the description of features combinations and resulting functions with clear meaning by a human data analyst, selecting the views and output interpretability. This approach differs from our work as we explore the combination of semantic types and high-level visual interpretation cues to classify data views representing relevant meaning.

2 Relevance Classification Model

2.1 Proposed Approach

The proposed relevance classification model consists of four main steps:

- Automatic pair-wise selection of attribute combination into a *data view*.
- Extraction of *descriptive statistical features* and *semantic type features*.
- Extraction of *visual interpretation cues*.
- Classification of the *relevance* of the data view.

Figure 1 presents an overview of the relevance classification model.

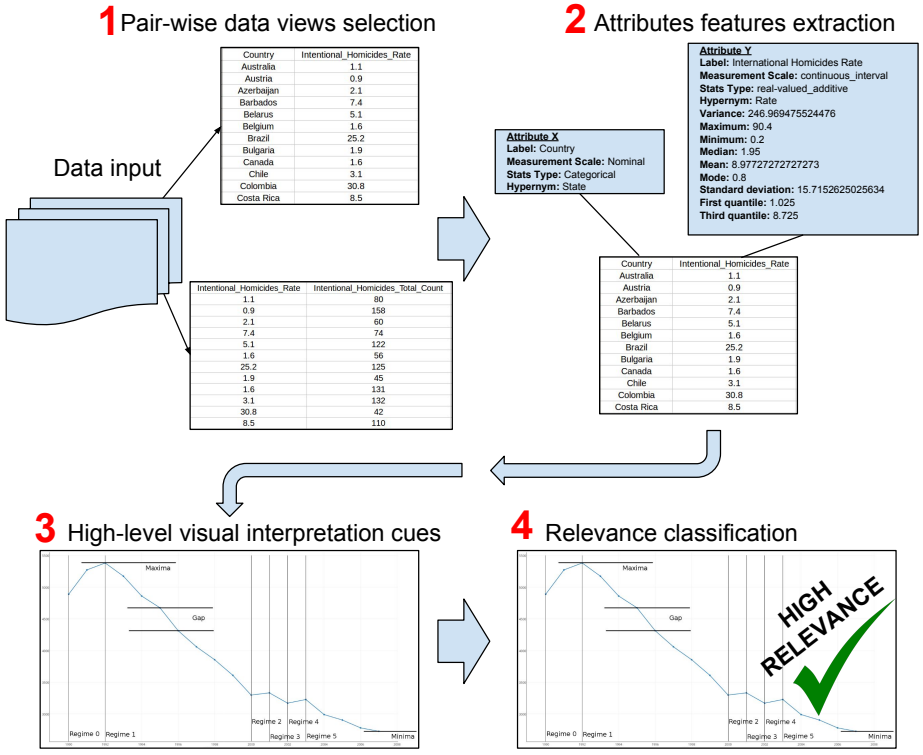

Fig. 1. Overview of our proposed approach.

We say that a data view is *relevant* when a visual interpretation provides a clear trend or pattern which is easily recognisable by a human, with or without previous knowledge about the data being analysed from two or more attributes in a dataset.

To classify the relevance when analysing plots of pairs of attributes (data views) we defined target classes based in the human process of data analysis. During the data analysis the analyst explore visualisations in order to understand the data. To achieve a clear meaning the analyst should take decisions, such as the application of operations (e.g. group by, sort by), changing the visualization plot type, and including more data attributes or dimensions. Thus, the classes defined are the representation of decisions required by the data analyst at each step of the exploration process. These classes are:

- Class 1 - Clear meaning - Generic (Very intuitive - you do not need to know the dataset/context to understand)
- Class 2 - Clear meaning - Dataset Context (you should know the dataset to understand)
- Class 3 - Data non-relevant for data analysis (or for the analysis in question)
- Class 4 - Label not equal to data semantics (Inconsistent data)
- Class 5 - Change visualisation (plot type or axis - makes sense, but if change it's better)
- Class 6 - Add operations (ex: group by, sort by, etc.)
- Class 7 - Additional data attributes needed for the interpretation
- Class 8 - Add operations and/or more data attributes and/or visualisation

Figure 2 shows an example of a plot that requires additional attributes to present a clear meaning. Figure 3 shows an example of a plot requiring an additional attribute, and/or an additional operation, and/or a change in the visualisation plot to present a clear meaning. Figure 4 shows an example of clear meaning.

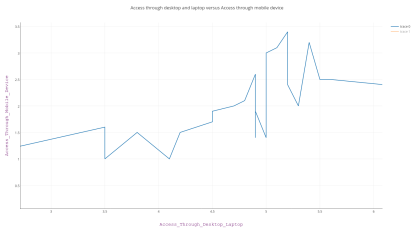

Fig. 2. Mobile devices and desktop/laptop devices by country sorted.

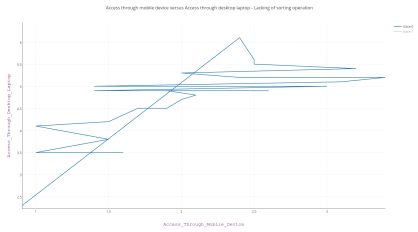

Fig. 3. Mobile devices and desktop/laptop devices by country without a sort operation.

In our approach we assume that the datasets have no missing values. Any missing values from the collected data is striped out before we start processing them.

2.2 Semantic Features

To represent the semantic type of an attribute, this work considers two approaches. The first one uses WordNet[1] hypernyms. WordNet is a lexical thesaurus for the English language, where words are organized into a lexical semantic network. WordNet also specifies the sets of hypernyms associated to a word (its taxonomical structure), where Y is a hypernym of X if every X is a (kind of) Y. We use these hypernyms as the semantic feature of words.

The other approach is the use of distributional vectors, extracted using the Word2vec framework (Mikolov *et al.* (2013)). These vectors encode co-occurrence statistics of words, relying on the linguistic notion that the context of a word defines the semantics of it (Harris (1954)). Distributional vectors are normally used as semantic representation of words.

For our work, we always assume that an attribute label has a descriptive meaning. For example, a label will never be 'X1' or 'Y'. This assumption is necessary if we want to automatically assign semantic features to a label.

2.3 Attribute Feature Extraction

In order to simulate the data analysis steps executed by humans, we developed a feature extraction process. The process explores the compositionality of statistical data types, the semantic representation of attributes, associated data operations and basic plotting resources.

For the extraction of *descriptive statistical features*, we consider the univariate analysis for description of the distribution, central tendency and the dispersion for each data attribute, also classifying the measurement scale and statistical data type. Examples of extracted features are presented in Table 1.

The set of semantic features are the *attribute labels*, the *WordNet hypernyms* of the labels and the *distributional vectors* of each label. Table 2 exemplifies some hypernyms used.

We end up with the following features:

- Statistical features:
  - Mean, median, first quartile, third quartile, mode (for categorical data), standard deviation, variance;
  - Measurement scale (nominal, ordinal, continuous interval, continuous ratio);
  - Statistical data type (categorical, ordinal, real, binary, multiclass, count);
- Semantic features:
  - Data labels;
  - WordNet hypernyms;
  - Distributional vector representations of data attributes labels;

The process of assigning hypernyms to the attributes consists in the identification of the head word of the phrase associated with the attribute label (when

[1] `http://wordnet.princeton.edu`

Table 1. Examples of extracted descriptive statistical features.

Attribute label	Measurement scale	Statistical Data Type	Median	Mean	First quart	Third quart	Mode	Std.Deviation	Variance
Access_Through_Mobile_Device	continuous_interval	real-valued_additive	2	2.125	1.575	2.525	1.9	0.689	0.475
SepalLength	continuous_interval	count	5.8	5.843	5.1	6.4	5	0.828	0.685
Intentional_Homicides_Total_Count	ordinal	count	88	89.53	48	130	28	47.18	2226.037
Intentional_Homicides_Year	ordinal	ordinal-integer_number	2012	2011.7	2012	2012	2012	0.569	0.324
Country	nominal	categorical	-	-	-	-	-	-	-
having_IP_Address	nominal	binary	1	0.314	-1	1	1	0.949	0.901
URL_Length	nominal	multiclass	-1	-0.633	-1	-1	-1	0.766	0.586
Price	continuous_ratio	real-valued_multiplicative	6.985	7.042	6.508	7.441	6.204	0.634	0.402
alcohol-use	continuous_ratio	count	2	2.176	0.5	77.5	49.3	26.878	722.473

Table 2. Examples of extracted semantic features.

Label	WordNet Hypernym
Access Through Mobile Device	activity
Country	area, social unit
Intentional Homicides Rate	rate
Intentional Homicides Total Count	count
Intentional Region	area
Intentional Homicides Year	time period
Homicides Total	total sum
Gun Homicides Sources and Notes	source, note
GDP Rank	rank
GDP Int.dolar	monetary unit

the label contains multiple words). A *word sense disambiguation* process selects the associated sense of the word considering the other words within the phrase as its context. Afterwards, the associated hypernym is assigned. The level of taxonomic abstraction is assigned to two taxonomic hops.

2.4 Visual Interpretation Cues

Another fundamental component of the proposed model consists in simulating the human visual interpretation process when analysing a data view (the combination of pairs of attributes).

The visual attention mechanism associated with the process of human data interpretation focuses on targeting the detection of outliers, coarse-grained variation regimes, clusters, periodicity, among others. These are examples of high-level visual features which provide an entry point to the interpretation of the data.

For the purpose of this approach, we identified a set of ten *high-level visual interpretation cues* described below. These cues are then used as features for our classifiers.

- Whether the function is a pair of numerical data or a pair containing at least one categorical data;
- Gaps;
- Quantity of existent gaps;
- Measure of numerical correlation;
- Whether the correlation is positive or negative;

– Whether the function is linear or nonlinear;
– Derivative regimes;
– Quantity of derivative regimes;
– Maxima/minima;
– Periodicity.

We define a gap as any considerable difference of value magnitude between consecutive data points. We call a "considerable difference" any value greater than the arithmetic mean of all the differences from consecutive or non consecutive data points.

Figure 4 show some high-level visual cues in the context of a data view.

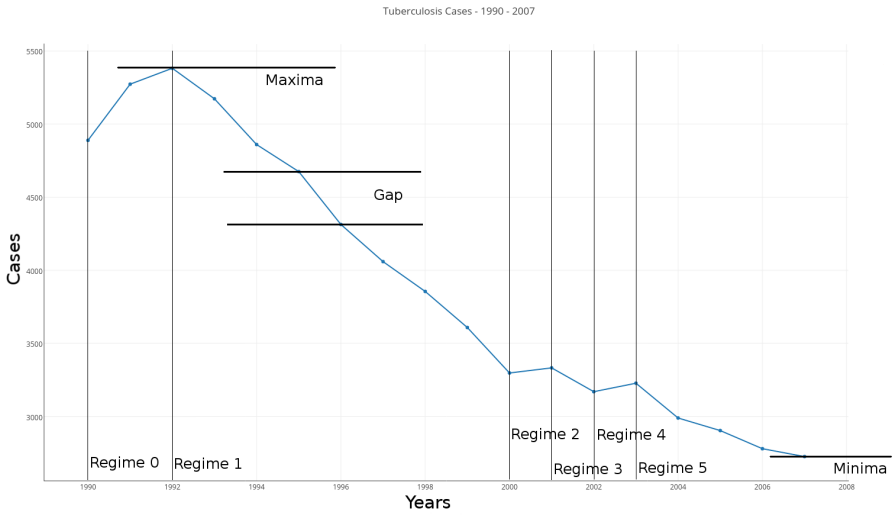

Fig. 4. Example of a two-dimensional numerical function describing the number of tuberculosis cases since 1990 up to 2007 with examples of visual interpretation cues.

3 Evaluation

Twenty-three machine learning models were trained to address the relevance classification problem, considering as an outcome one of the eight classes previously presented. We applied machine learning models based on *(i)* linear, *(ii)* non-linear, *(iii)* non-linear with decision trees, *(iv)* non-linear with boosting and *(v)* neural networks approaches. The application of more than one machine learning model is intended to assess the behaviour of our dataset and evaluate the impact of semantic and visual features in the predictive modelling.

3.1 Creation of the Relevance Gold-standard

Our gold-standard dataset consists of 20 open datasets commonly used for data analysis and machine learning tasks[2], available at the UCI repository[3], Plotly[4], EDX Analytics Edge[5], and William B. King R Tutorials[6]. All attribute pairs from the collected datasets are then plotted and classified as one of the eight classes already presented previously, resulting in 2989 attribute pairs.

Two types of visualization plots were applied, the *bar plot* and *scatterplot with lines*, following the rules: *(i) bar plot* is applied when at least one data type is qualitative and *(ii) scatterplot with lines* is applied when the X axis is a quantitative data type and Y axis is a quantitative data type. Based on the plotting, a human classified the relevance class of each attribute pair.

3.2 Experiments

For the training of the machine learning models, we represented the extracted features in five scenarios:

- using only statistical features;
- using visual interpretation cues;
- using WordNet hypernym and visual interpretation cues;
- using distributional vectors of the attribute labels and visual interpretation cues;
- using distributional vectors of random words and visual interpretation cues.

We use the last two scenarios to validate our assumption that semantic vectors of the labels would improve the predictability accuracy of classification. The variation of the semantic types is used to evaluate the impact of different semantic features representations.

In all of the experiments we used 80% of our dataset instances in the training phase, using the remaining 20% to validate the resulting models.

3.3 Results

The best classification results are achieved with the feature combination of hypernyms as semantic types and visual interpretation cues (Random Forest achieves a 11.87% improvement in F1 score over the best result of the scenario without these features). Distributional semantic vectors also impact in the classification of the results (5.1% improvement in F1 score over the best result of the scenario without distributional semantic vectors). The best F1 score (0.906) was achieved

[2] `https://github.com/ekamioka/unipassau-ada`
[3] `http://archive.ics.uci.edu/ml/`
[4] `https://plot.ly/`
[5] `https://www.edx.org/course/analytics-edge-mitx-15-071x-2#!`
[6] `http://ww2.coastal.edu/kingw/statistics/R-tutorials/multregr.html`

Table 3. Results of the relevance classification. Best AUC and F1 score for each classifier is highlighted. HypVis - With hypernyms and visual interpretation cues, Vis - Just with visual interpretation cues, WO - No semantic features and no visual interpretation cue, VecVis - With distributional representations of the labels and visual interpretation cues, VecrVis - With distributional representations of random words and visual interpretation cues. The last line shows the percentual improvement for each feature set.

	HypVis		Vis		WO		VecVis		VecrVis	
Algorithm	AUC	F1 Score	AUC	F1 Score	AUC	F1 Score	AUC	F1 Score	AUC	F1 Score
Latent Discriminant Analysis	**0.742**	0.658	0.664	0.578	0.469	0.323	0.735	**0.667**	0.659	0.531
Linear Support Vector Classification	**0.455**	0.430	0.398	0.517	0.399	0.502	0.384	0.536	0.283	**0.554**
Stochastic Gradient Descent	**0.522**	0.035	0.478	0.055	0.466	0.065	0.518	0.014	0.394	0.017
SGDClassifier with kernel approximation	**0.496**	0.380	0.470	0.385	0.423	**0.460**	0.492	0.373	0.485	0.344
R Kernel Support Vector Machine	**0.868**	**0.754**	0.687	0.622	0.811	0.656	0.807	0.623	-	-
Naive Bayes	0.647	**0.617**	0.671	0.543	0.632	0.545	**0.680**	0.530	0.644	0.500
R k-Nearest Neighbors 3 (k=5)	0.798	0.698	0.723	0.603	0.687	0.426	0.782	0.610	**0.866**	**0.767**
Sklearn k-Nearest Neighbors (k-10)	0.320	**0.680**	0.314	0.665	**0.330**	0.645	0.320	**0.680**	0.320	**0.680**
R Classification and Regression Trees(CART)	0.793	0.620	0.784	0.629	0.713	0.571	0.651	0.538	**0.797**	**0.652**
Sklearn DecisionTreeClassifier	0.272	**0.763**	0.276	0.741	**0.301**	0.751	0.273	0.762	0.330	0.688
ExtraTreesClassifier	0.282	0.752	0.306	0.699	0.322	0.651	0.268	**0.762**	**0.326**	0.534
C4.5 Weka	0.633	0.074	0.658	0.074	0.729	0.667	**0.873**	**0.825**	0.679	0.082
PART Weka	0.688	0.540	0.892	**0.738**	0.804	0.657	**0.905**	0.702	0.510	0.346
R Random Forest	**0.969**	**0.895**	0.882	0.765	0.915	0.779	0.919	0.794	0.781	0.667
Sklearn RandomForestClassifier	0.263	**0.792**	0.282	0.762	0.287	0.757	0.263	0.775	**0.349**	0.614
R Gradient Boosted Machine	**0.646**	0.375	0.543	**0.757**	0.551	0.425	0.577	0.170	0.625	0.318
R Boosted C5.0	0.904	0.767	0.932	0.862	0.899	0.800	0.925	**0.906**	**0.951**	0.875
R eXtreme Gradient Boosting	0.916	0.759	0.914	0.783	0.878	0.772	0.925	**0.851**	**0.926**	0.818
Sklearn AdaBoostClassifier with: 10 max_depth	0.278	0.780	0.275	0.778	0.288	0.777	0.261	**0.791**	**0.340**	0.667
R Simple Neural Networks - nnet	0.559	**0.757**	**0.759**	0.729	0.716	0.518	-	-	-	-
H20 Deep Learning	0.958	**0.867**	0.944	0.741	**0.968**	0.731	0.875	0.642	-	-
Multi Layer Perceptron - tanh	**0.538**	0.358	0.536	0.076	0.457	0.431	0.341	**0.490**	0.508	0.237
Multi Layer Perceptron - relu	0.452	0.428	0.362	**0.485**	0.522	0.020	0.483	0.394	**0.500**	0.376
Wins percentage	**34.78%**	**34.78%**	4.35%	13.04%	13.04%	8.69%	13.04%	**34.78%**	**34.78%**	17.39%

by using a Boosted C5.0 classifier, using distributional vectors of attributes labels and visual interpretation cues. The full comparative analysis of different classification methods and features are fully presented in Table 3.

Considering the imbalanced problem in the classification dataset, we noted a better classification performance in nonlinear models and ensemble-based models, which implements resampling techniques and combinations, thus rebalancing the classes at learning time (Chawla (2005)).

Other classifiers that do not implement some type of rebalancing have a hard time classifying some instances. For instance, the samples labeled as *Class 7*, a class that has only 7 occurrences in our dataset, are rarely classified correctly. On the other hand, our most common class (*Class 2*), represents 52.12% of our dataset.

To further interpret our classifier results we hand-picked some examples. Those examples are presented in figures 5, 6 and 7.

4 Conclusion and Future Work

This work proposes a classification model for data relevance using the combination of *semantic types* and *high-level visual interpretation cues*. After performing a systematic comparative analysis of different classifiers, the proposed model

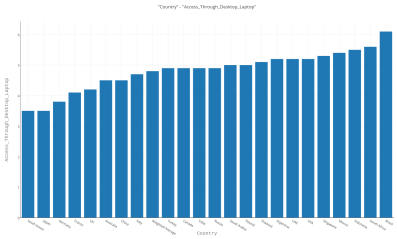

Fig. 5. Country by Access_Through_Desktop_Laptop. Correct classification of Class 1 (Clear meaning and very intuitive).

Fig. 6. CPI_all by Unemployment in a dataset about Elantra Sales. Correct classification of Class 2 (Clear meaning dependent of dataset context).

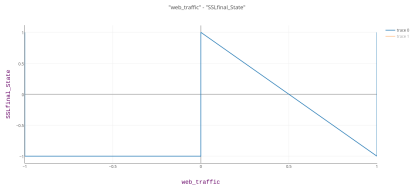

Fig. 7. website_traffic by SSLfinal_state in a dataset about detection of phishing attack in webpages. Correct classification of Class 8 (Requires additional operations and/or data attributes/dimensions and/or different plot type to depict a clear meaning).

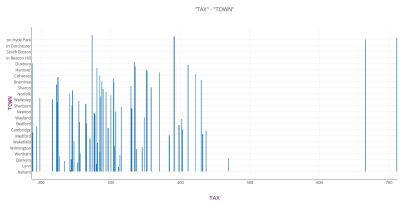

Fig. 8. TAX by TOWN in dataset about Boston price location. Class 5 (Change visualisation) misclassified as Class 2 (Clear meaning dependent of dataset context).

achieves a 0.906 F1 score using a Boosted C5.0 classifier, using distributional vectors of attributes labels and visual interpretation cues. The relevance classification model can be used to classify relevance of new data views. Additionally, the evaluation shows that semantic features and visual interpretation cues have a clear impact on classification performance.

Our approach currently does not cover use cases where missing values are present. The treating of missing values is crucial for real-world applications. This limitation should be addressed in future work.

Also, in practice, relevant patterns can be found in higher than 2-dimensional data views. We intend to apply the same proposed approach to higher dimensional data views and analyze the results.

Bibliography

Botia, J. A., Garijo, M., Bot'ia, J., Velasco, J. et Skarmeta, A. (1998). A generic datamining system. basic design and implementation guidelines.

Bremm, S., von Landesberger, T., Bernard, J. et Schreck, T. (2011). Assisted descriptor selection based on visual comparative data analysis. *In Computer Graphics Forum*, volume 30, pages 891–900. Wiley Online Library.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. *In Data mining and knowledge discovery handbook*, pages 853–867. Springer.

de Souza, D. F. P. (2015). *TIME-SERIES CLASSIFICATION WITH KERNELCANVAS AND WISARD*. Thèse de doctorat, Universidade Federal do Rio de Janeiro.

Dinsmore, T. W. (2014). Automated predictive modelling. [Online; posted 09-April-2014].

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B. et Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*.

Grosse, R., Salakhutdinov, R. R., Freeman, W. T. et Tenenbaum, J. B. (2012). Exploiting compositionality to explore a large space of model structures. *arXiv preprint arXiv:1210.4856*.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B. et Ghahramani, Z. (2014). Automatic construction and natural-language description of nonparametric regression models. *arXiv preprint arXiv:1402.4304*.

Lubinsky, D. et Pregibon, D. (1988). Data analysis as search. *Journal of Econometrics*, 38(1-2):247–268.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. et Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Spott, M. et Nauck, D. (2006). Towards the automation of intelligent data analysis. *Applied Soft Computing*, 6(4):348–356.

St Amant, R. et Cohen, P. R. (1997). Interaction with a mixed-initiative system for exploratory data analysis. *In Proceedings of the 2nd international conference on Intelligent user interfaces*, pages 15–22. ACM.

St. Amant, R. et Cohen, P. R. (1998). Intelligent support for exploratory data analysis. *Journal of Computational and Graphical Statistics*, 7(4):545–558.

Záková, M., Křemen, P., Železný, F. et Lavrač, N. (2011). Automating knowledge discovery workflow composition through ontology-based planning. *Automation Science and Engineering, IEEE Transactions on*, 8(2):253–264.