

FDive: Learning Relevance Models using Pattern-based Similarity Measures

Frederik L. Dennig¹, Tom Polk¹, Zudi Lin², Tobias Schreck³, Hanspeter Pfister², and Michael Behrisch²

¹University of Konstanz, Germany*

²Harvard University, USA[†]

³Graz University of Technology, Austria[‡]

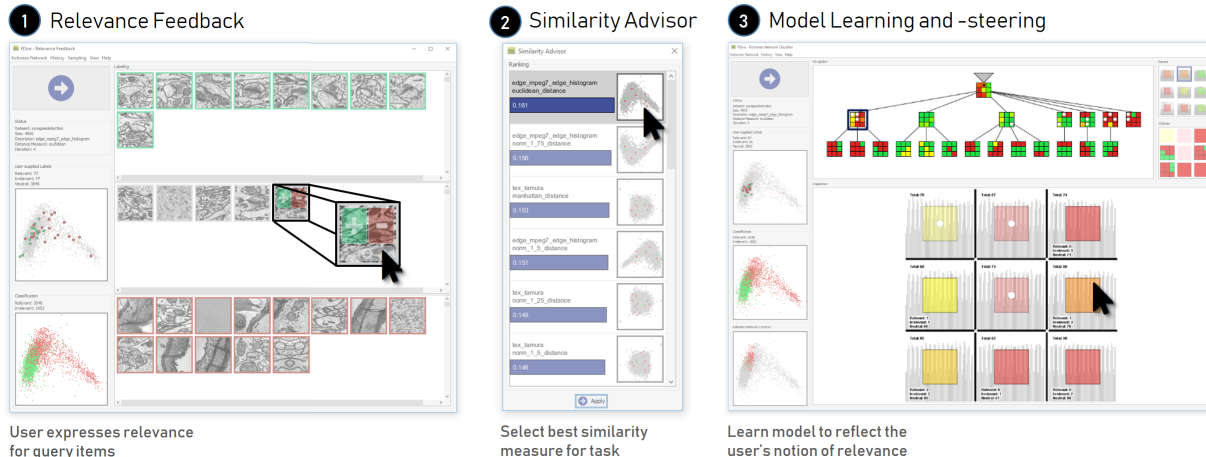


Figure 1: FDIVE learns to distinguish relevant from irrelevant data through an iteratively improving classification model by learning the best-fitting feature descriptor and distance function. (1) Users express their notion of relevance by labeling a set of query items, in this case, images. (2) These labels are used to rank all similarity measures by their ability to distinguish relevant from irrelevant data. (3) The system applies the selected similarity measure to learn a Self-Organizing Map (SOM)-based relevance model. Users explore and refine the model by supplying relevance labels in uncertain data regions, especially near the decision boundaries.

ABSTRACT

The detection of interesting patterns in large high-dimensional datasets is difficult because of their dimensionality and pattern complexity. Therefore, analysts require automated support for the extraction of relevant patterns. In this paper, we present FDIVE, a visual active learning system that helps to create visually explorable relevance models, assisted by learning a pattern-based similarity. We use a small set of user-provided labels to rank similarity measures, consisting of feature descriptor and distance function combinations, by their ability to distinguish relevant from irrelevant data. Based on the best-ranked similarity measure, the system calculates an interactive Self-Organizing Map-based relevance model, which classifies data according to the cluster affiliation. It also automatically prompts further relevance feedback to improve its accuracy. Uncertain areas, especially near the decision boundaries, are highlighted and can be refined by the user. We evaluate our approach by comparison to state-of-the-art feature selection techniques and demonstrate the usefulness of our approach by a case study classifying electron microscopy images of brain cells. The results show that FDIVE enhances both the quality and understanding of relevance models and can thus lead to new insights for brain research.

Keywords: Visual analytics, similarity measure selection, relevance feedback, active learning, self-organizing maps.

*e-mail: {frederik.dennig, thomas.polk}@uni-konstanz.de

[†]e-mail: {linzudi, pfister, behrisch}@seas.harvard.edu.

[‡]e-mail: tobias.schreck@cg.v.tugraz.at

1 INTRODUCTION

A primary challenge when analyzing collected data is to distinguish relevant from irrelevant data items. Large and high-dimensional datasets are not easily analyzed, because of their size, dimensionality, and possible complex patterns. Therefore, analysts need automated support. This support is realized in the form of a relevance model that can help them to make this distinction. Its task is the retrieval of relevant data items from large high-dimensional datasets that are often associated with many types of analysis scenarios. Similarity models are key to effective data clustering and classification. It is crucial that the model reflects the notion of relevance as it pertains to the analysis task. More generally, when we are dealing with high-dimensional datasets, we need to automatically and adaptively assess the relevance of data items. Although analysts interact with data for analysis and exploration purposes, their primary goal is to quickly generate new insights and results. All interactions, such as labeling or relevance feedback, should be focused on yielding insights and need to be as impactful as possible.

The fully automatic creation of relevance models is non-trivial. Deep learning approaches, such as Convolutional Neural Networks (CNNs), have been applied successfully, but typically require a large number of labeled training data to distinguish relevant from irrelevant data [38]. Classic machine learning techniques depend on a predefined set of features and a given distance function, chosen or even designed by experts based on their experience. In most real-world scenarios, these labels do not exist and the manual assignment of labels is time consuming, tedious, and expensive. In many analysis scenarios, this is not a viable solution. Transfer learning could be an alternative solution. These methods reapply a previously learned model for a different task than that for which they were originally trained [48]. While the idea seems intriguing, these models are unable to transfer the complex user understanding between datasets.

One reason is that the problem and task definition in exploratory scenarios, particularly the pattern space, is highly specific and non-static. Users' mental model of *what makes up relevance* evolves throughout an analysis, thus requiring adaptive methods for the process. Additionally, the created model needs to be understandable, explorable, and refinable in areas where it is inaccurate.

The feedback-driven view exploration pipeline by Behrisch et al. [5] was an early approach towards a relevance model-guided exploration of large multidimensional datasets. Similar to our work, the central idea is to make an arbitrary dataset accessible to users through visualizations, such as scatter plots, that can be abstracted into a set of numbers, called features. To solve real-world problems, features need to be able to express the differences between the data items concerning the analysis objective. Features reduce the complexity of comparing data items but are limited in their ability to express all properties of a data item. The approach by Behrisch et al. [5] only uses one *fixed* feature descriptor (FD), namely Scagnostics [63], limiting the set of described properties and introducing biases into the analysis process. In this work, we tackle the question of choosing an appropriate FD that models the given dataset, analysis domain, and analysis task. We claim that FDs alone do not express the relationship between data items. We also need a distance function that describes their relationships. Depending on the analysis scenario, other measures than the ubiquitous Euclidean distance may perform significantly better [22], which reflects on the performance of the relevance model learning component, too. In this work, we expand on Behrisch et al.'s static decision tree model, in which exploration decisions are irreversible, with a more flexible and adaptive approach to guide the user through the data space. Our classification results and feature abstraction can be visually explained, making the quality of the model easier to capture and more trustworthy.

In this work, we present FDIVE, a visual analytics system for the creation of relevance models. In FDIVE, we model relevance as a binary classification problem. Since the quality of the underlying classification or ranking model depends on the usefulness of the employed FDs and distance function, we introduce the concept of the *Similarity Advisor* engine, which ranks FD-distance function pairs, according to their ability to distinguish relevant from irrelevant data. This removes the need for an expert choosing an FD and distance function manually. The system uses the best-ranked similarity measure for the creation of the relevance model. To learn fine-grained differences between relevant and irrelevant data, we introduce a Self-Organizing Map (SOM)-based relevance model that classifies data items according to their cluster membership. To allow the judgment of the model quality and model refinement, the SOM-based model is visually explorable and guides the user towards areas of uncertainty. We embed the *Similarity Advisor* and the model learning process into an iterative framework, to allow for convergence towards the optimal similarity measure and relevance model.

We evaluate our general framework through a quantitative study comparing FDIVE to three state-of-the-art feature selection techniques, where we show that the *Similarity Advisor* can outperform them in scenarios with a low number of labels through a fast adaptation to the user's notion of relevance. We also demonstrate FDIVE's applicability and usefulness on a challenging scientific analysis task. Specifically, we consider electron microscopy images of brain cells, where a domain expert teaches the system the relevance of images depicting a neuronal synapse.

2 RELATED WORK

In this section, we delineate FDIVE from other approaches. FDIVE is a relevance model builder, in contrast to image retrieval systems like PixSearcher [47] which enables users to retrieve images through query by example. In the following, we discuss related concepts such as feature selection, visual active learning, and distance function

learning. We also discuss similarities and differences in the area of model visualization and understanding.

2.1 Feature Selection for Dimensionality Reduction

Feature selection algorithms typically try to approximate the usefulness of a given feature. These techniques determine a subset of relevant feature dimensions based on feature-ranking and feature-weighting [23, 31]. Although prior studies show how visualizations can support feature selection and optimization in 3D models [56] or exploration of chemical compounds [10, 57], the feature evaluation procedure is reoccurring and potentially exhausting for the user. Thus, we decided to use two purely automatic statistical feature selection algorithms in the evaluation of FDIVE. First, *ReliefF* [37, 62] is a state-of-the-art extension of the *Relief* algorithm for multi-class problems [44]. It ranks features based on how well they distinguish an instance from its k -nearest neighbors. If a neighbor is from a different class, the weights of features that separate both instances are increased, and all others are decreased accordingly. In case the neighbor is from the same class, the weights of features that separate both instances are decreased, and all others increased. Second, *Linear Ranking Ensembles* combine multiple ranking classifiers, such as the *Recursive Elimination Support Vector Machine (SVM)*, into one ranking ensemble. They are, thus, more stable than other approaches [55]. *Recursive Elimination SVMs* iteratively reduce the feature dimensions size using linear SVMs [40]. Attributes are ranked, and the worst performing dimension is removed. This process, including the SVM training, continues until only one feature dimension remains.

The quality of a feature selection depends on the number of available labels and is computationally expensive in scenarios that require continuous reevaluation. With FDIVE, we provide a solution for this scenario by keeping the feature descriptions while ranking a set of similarity measures, consisting of an FD and a distance function combination, based on how well it separates relevant from irrelevant data. We embed this technique in an iterative process, allowing for an adaptation to the best-suited similarity measure.

2.2 Visual Active and Interactive Machine Learning

In a visual active learning (AL) system, users are provided with auxiliary information about the learning process and model state, specifically decision boundaries of the classification model, query choice, and learned instances. Bernard et al. [8] present a visual AL method to assess the well-being of prostate cancer patients from the patient's history, describing interesting biological and therapy events. The tool suggests a set of candidates to label, as well as allowing for the visual verification of the validity of learned instances. Heimerl et al. [28] present a visual AL system as an SVM classifier for text. The tool supports the visualization of the decision boundary, including instances on it, and user-based instance selection for labeling. Eaton et al. [19] adjust the underlying data space by describing it with manifold geometry, allowing users to label data items, serving as control points leading to improved learning performance.

In contrast to AL, the sample selection in interactive machine learning (IML) is driven by the user. Dudley et al. [17] describe a general approach to interface design for IML providing an overview of challenges and common guiding principles. Arendt et al. [2] present an IML interface with model feedback after every interaction by updating the items shown for each class. The users can drag misplace data items to the appropriate class and, if needed, create a new one. Both actions update and improve the model.

FDIVE is a visual active learning system that learns a relevance model based on the user's notion of relevance. We propose a SOM-based model, which is interactively explorable, guiding the user to areas of uncertainty and decision boundaries. The model creation and inspection are combined in an iterative workflow that allows

the user to observe and judge model change, leading to a more understandable relevance model and learning process.

2.3 Distance Function Learning

Another requirement to represent the relationship of data items is a distance function. A distance function can include a feature weighting. The Mahalanobis metric [43] measures the standardized distance of a data point to the estimated mean of its population. Relevant Component Analysis [3] uses a parameterized Mahalanobis distance. This technique adapts the feature space by assigning large weights to relevant dimensions and low weights to irrelevant dimensions through equivalency constraints, describing the similarity of data items. As opposed to purely algorithmic approaches, there are also visual and interactive approaches to the generation of suitable distance functions. Brown et al. [11] learn a distance function from a 2-dimensional projection of the data space where the user drags the data point to the desired position, thus describing similarity relations. The underlying distance function is updated accordingly by the adaptation of feature weights. The work by Gleicher [21] demonstrates the learning of multiple distance functions, each describing the relationship of the data based on different features, capable of describing abstract concepts, such as socio-cultural properties of cities. Fogarty et al. [20] present an image retrieval system that determines the weights of a distance metric based on user-supplied feedback to learn concepts.

In contrast, FDIVE unifies many concepts mentioned above. It ranks arbitrary feature descriptors and similarity measure combinations by their ability to discriminate relevant from irrelevant data. FDIVE removes the limitation on a pre-defined set of features through the comparison of multiple FDs describing a diverse set of data properties. Also, a set of similarity coefficients is used, thus removing the limitation of a single similarity coefficient or feature weighting. This makes FDIVE a generalized relevance model builder for different types of data.

2.4 Model Visualization and Understanding

Visual Analytics (VA) aims to provide the analyst with visual user interfaces that tightly integrate automatically obtained results with user feedback [34]. The knowledge generation model [54] describes an iterative process of exploration and verification activities of both human and machine. Results are presented visually to analysts, who interpret obtained patterns and provide feedback to steer the exploration process or form and refine hypotheses. The understanding and interpretation of machine-learned models is key for the effective incorporation of user feedback in such scenarios. Several prior works have studied model visualizations and interactions. BaobabView [61] presents a model where the structure of a decision tree is augmented with data distributions and data flows. Liu and Salvendy [41] and Ankerst et al. [1] use icicle plots [35, 39] to visualize decision trees. Visual interactive approaches for cluster evaluation and understanding were presented by Nam et al. for general high-dimensional data [46] and by Ruppert et al. [52] for the clustering of text documents. Sacha et al. present SOMFlow [53], an exploration system that uses Self-Organizing Maps (SOM) to guide the user through an iterative cluster refinement task, leveraging the proximity-preserving property of SOMs [7, 59] for clustering and data partitioning tasks.

In a model creation task, the user needs to be guided towards areas of high uncertainty. Thus FDIVE steers the data exploration to specific parts of the model, such as the decision boundaries. The SOM-based model of FDIVE is capable of providing the necessary information about uncertain areas and automatic refinement.

3 SIMILARITY ADVISED MODEL LEARNING

The key idea of our approach is to iteratively and interactively create relevance models, where a useful feature description is unknown,

and no or only few labels are available. Our proposed *Similarity Advisor* allows approaching the question which feature descriptor and similarity measure combination is useful to distinguish relevant from irrelevant data items. In a scenario where labels are sparse, the quantitative validation of classification models with performance measures is inexpressive. Thus, there is a need for techniques that allow for model assessment without test data. Classifiers, such as SVMs, have been used in visual active learning approaches [28]. However, the representation of the data space created by SVMs does not allow the user to judge the quality of a classifier visually. Decision trees are more intuitively interpretable.

We propose a SOM-based classification model which is embedded in an iterative workflow to allow for observable learning steps. In each step, the model is explorable and refinable to judge and improve its quality. Both, the *Similarity Advisor* and the SOM-based classification model constitute FDIVE, a generalized model builder. In the following, we provide an introduction to SOMs.

Self-Organizing Maps: FDIVE relies on a *neural network architecture*, called Self-Organizing-Map (SOM) or Kohonen Network. SOMs are the basic building block of our relevance model and are one of the classical neural network structures, created by Kohonen to derive topologically coherent feature maps [36]. SOMs can be visualized as a grid of cells representing the neurons of the network. The cells contain prototype vectors representing data clusters. In the learning phase of the network, the most similar prototype vector (best-matching-unit) to the training input is identified and adjusted towards the input vector. Spatially close neighbors are also adapted, depending on a learning rate and radius parameter. The latter gives rise to the self-organization property of the map. The final result is a topology, where data items are clustered. Clusters can consist of single or multiple cells, and cluster similarity can be captured by spatial proximity of clusters on the SOM grid [7, 53].

We extend this algorithm into a tree-like classifier to allow for the representation of fine-grained similarity differences. This concept is based on the idea that items can “flow” from a parent SOM node into a child SOM for further analysis, as presented by Sacha et al. [53]. In our work, we extend this idea to create a classification model that *automatically* partitions the high-dimensional data space into relevant and irrelevant data item clusters. We will detail this approach in Sec. 6. We use an interactive SOM visualization to allow for the visual inspection of the currently learned model, e.g., where groups of relevant or irrelevant data elements are located, and how well decision boundaries can distinguish known groups.

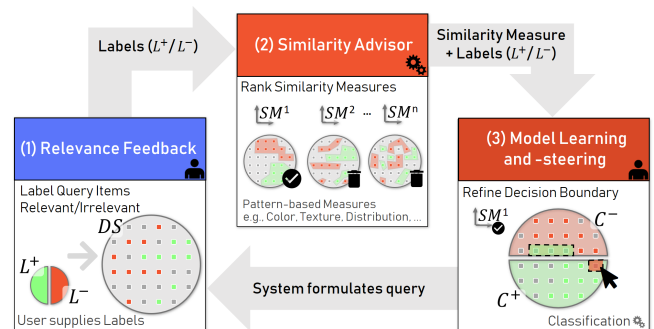


Figure 2: (1) Users label query items as relevant or irrelevant and therein express their notion of relevance. (2) This selection is used to automatically determine the best-fitting similarity measure, which distinguishes relevant from irrelevant data. (3) The system adapts the model using the relevance labels and similarity measure. The model is explorable and refinable by the users, to improve its accuracy.

3.1 Workflow for Iterative Relevance Model Learning

FDIVE is inspired by the feedback-driven interactive exploration tool by Behrisch et al. [5], which propose an iterative and FD-based exploration framework. A central principle is to represent an arbitrary dataset with the help of visualizations to make it accessible for an analyst. This visualization needs to be translated into a language understood by a computer, which uses this *proxy* to guide through the information- and pattern space which is achieved by a single fixed FD introducing bias into the analysis process.

We expand this body of work by changing the focus from an exploratory approach to a model building technique. The validation of relevance models, though, is a challenging task, due to the following reasons. We need to define a useful definition of similarity, but a metric for separating classes can only be determined during the learning process. What is needed are flexible and adaptive strategies for determining a useful metric defining similarity. FDIVE allows for arbitrary data modeling through the *Similarity Advisor*, which ranks a set of FDs and distance functions by their usefulness concerning the current analysis domain and dataset properties. The FD, representing the data modeling, is subsequently used to create a relevance model. Additionally, the model needs to be explorable and refinable to convince an analyst of its usefulness and accuracy.

In FDIVE, we leverage an iterative workflow to continuously revalidate the similarity measure and improve the relevance model. In the following, we describe each iteration step and its impact. Fig. 2 shows each step accordingly.

(1) Relevance Feedback: The system prompts the user to label a subset of data items of the dataset (DS) as relevant or irrelevant, representing relevance as a binary classification problem. Those data items labeled as relevant are referred to as \mathcal{L}^+ and all labeled as irrelevant as \mathcal{L}^- . Unlabeled data items are considered neutral. In the first iterations, this step is replaced by a query generated through a representative data sample. In all following iterations, the query is determined by the SOM-based model. FDIVE supports the user by visual feedback allowing the validity assessment of a currently used similarity measure and classification through visual feedback. This step is described in detail in Sec. 4.

(2) Similarity Advisor: The system evaluates all possible pairwise combinations of FDs and distance functions by their ability to separate relevant (\mathcal{L}^+) from irrelevant (\mathcal{L}^-) data items. A ranking shows the evaluation result, giving an intuition about the similarity measures. The user can follow the recommendation or choose a different similarity measure. The system uses the FD and distance function for the creation of the relevance model. We describe the algorithmic background of the *Similarity Advisor* in Sec. 5.

(3) Model Learning and Steering: The system creates a classification model based on the selected similarity measure and available labeled data (\mathcal{L}^+ and \mathcal{L}^-). The model can be explored to assess its properties and viability for its classification task. The classification result is referred to with \mathcal{C}^+ describing all data items classified as relevant and all irrelevant as \mathcal{C}^- . The SOM-model creation and interactions are described in Sec. 6. Subsequently, the system determines a set of query items which are labeled by the user in the first step of the next iteration.

In the following, we describe the design, user interaction and algorithmic support in FDIVE.

4 CONTEXT-AWARE RELEVANCE FEEDBACK

Data labeling is the first and reoccurring step in our relevance model learning process from Sec. 3.1. During start-up, this essential bootstrapping step helps us to form a decision basis for the subsequent application of our *Similarity Advisor*. Throughout the learning process the classifier queries relevance labels through this interface to improve its accuracy. We describe this step of FDIVE in Sec. 6.

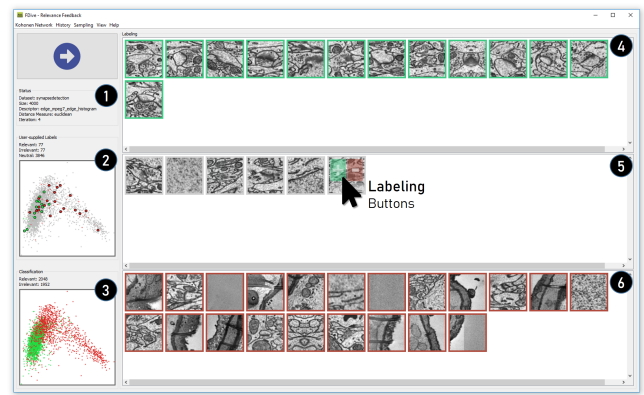


Figure 3: Context-aware Relevance Feedback: (1) Status display showing the current analysis state. (2) Scatter plot highlighting newly labeled data. (3) Scatter plot of the current classification result. Both allow judging the impact of new labels. (5) Queried neutral data items. (4) Data items labeled as relevant and irrelevant (6).

4.1 Relevance Feedback of Representatives

We sample data items in the first iteration for an initial user labeling. The sampling method can be chosen from the following options: Minimum-Maximum-, Quantile Sampling, Normal-, Stratified Normal Bootstrapping, Normal- or Stratified Subsampling [5]. In all following iterations, the request for labeling is determined by the relevance model, in our case a Self-Organizing Map-based model (Sec. 6). The user can apply three types of labels: relevant, irrelevant and neutral. While the relevant and irrelevant labels express a user preference and have an impact on all steps of FDIVE, neutral represents an uncertain item. The model may prompt a label for the given element at a later iteration. The user labels a subset of displayed data items by clicking on the mouse-over menu or using a keyboard-shortcut. For visual clarity, all elements are assigned to specific panels (relevant, neutral, irrelevant, from top to bottom in Fig. 3 (3-5), according to their label type, which also allows comparing items with the same relevance label.

4.2 Visual Assessment of Labeling Impact

A status display (Fig. 3 (1)) provides information about the current analysis state, such as the current FD and distance function, the number of supplied relevant and irrelevant labels, as well as the number of remaining neutral items. A scatter plot (Fig. 3 (2)) of the dataset using the currently chosen FD and distance function depicts the possible impact of new labels when compared to the projection of the classification result (Fig. 3 (3)). We create both 2D projections with multi-dimensional scaling (MDS). MDS projects the data in a distance-preserving way without the need for additional parameters. The annotation view is also used to refine the labels in the SOM-based model and explore elements assigned to a SOM-neuron (Sec. 6). Chegini et al. explored the idea of showing the classification result in a scatter plot [14], while the visual feedback on data labeling was evaluated by Bernard et al. [6]. Combining both approaches allows assessing the impact of newly assigned labels in a natural form. The comparison of both scatter plots shows the effect of new labels, e.g., a relevant label in an area of irrelevant classifications hints at an incomplete reflection of the user’s notion of relevance, a matching label hints at a convergence.

5 ASSESSING PATTERN-BASED SIMILARITY MEASURES

The goal of the *Similarity Advisor* is to select the most expressive FD and distance function combination from a predefined set of FDs and distance measures to improve the relevance model creation. We

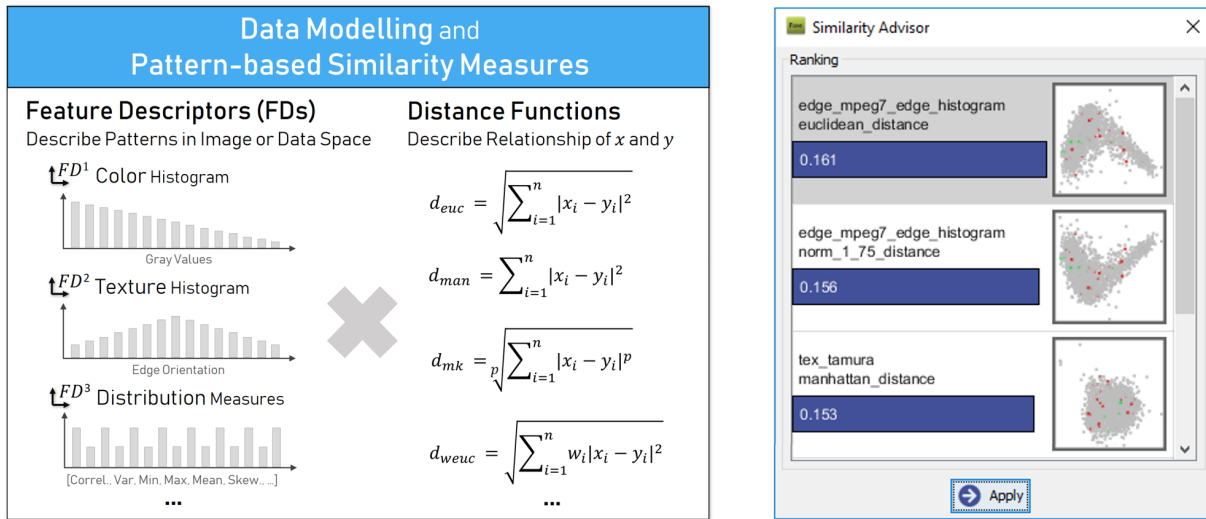


Figure 4: Left – The *Similarity Advisor* uses a set of FDs and distance functions. FDs model the data based on perceptible patterns in the data or image space. Distance functions describe the relationship between two points in the FD space. In FDIVE, we consider all pair-wise combinations as potentially useful measures. We call a combination of an FD and a distance function a *pattern-based similarity measure*. Right – The *Similarity Advisor* ranks all pair-wise combinations of FDs and distance functions according to their ability to distinguish relevant from irrelevant data. A bar indicates the score and a scatter plot shows the topology of implied data distribution allowing users to judge its usefulness.

claim that a combination of FD and distance measure can define a *pattern-based similarity measure*. To describe the discriminative ability, we need a *quality metric* that reflects the similarity measure’s ability to distinguish between our relevant and irrelevant items. We consider a useful similarity measure one that maximizes the distance between both sets \mathcal{L}^+ and \mathcal{L}^- . We considered other quality metrics, such as metrics that measure distances between elements of a cluster, but found them lacking in performance. We propose the *Similarity Advisor* for the selection of a suitable distance metric; this includes the choice of an FD and a distance function. For this, we require a set of diverse FDs. We use various FDs from the Image Processing and Computer Vision Community because these algorithms are designed to match the human perceptual system.

In essence, the application scenario determines the usefulness of a feature description and distance function. However, the selection of a useful distance function is hard. Thus, we introduce the concept of continuously evaluating a set of *pattern-based similarity measures* for their applicability to the current analysis task, allowing for the convergence to the most useful one. To describe the algorithmic basis of the *Similarity Advisor*, we define all relevant terms.

Feature Descriptor (FD): FDs are modeling specific characteristics of a data item. Examples for low-level FDs are color histogram descriptors, modeling the color distribution, or edge histograms describing edge orientations of an image [51]. Low-level FDs are typically inexpensive to compute and may work robustly. Depending on the type of data at hand, many FDs are applicable. Mathematically, an FD can be described as a function $FD : DS \rightarrow \mathbb{R}^n$, where DS denotes the dataset and \mathbb{R}^n the implied vector space. The dimensionality n depends on the FD. Table 1 lists all FDs used by FDIVE. These FDs describe a variety of different image features, such as color, layout, structure, and shape [4].

Feature Vector (FV): An FV is an instantiation of an FD for a specific data item. An FV contains one or multiple components, called feature dimensions or features. A feature vector $FD(x) \in \mathbb{R}^n$ represents a description of a data item $x \in DS$, w.r.t. the properties described by the applied FD.

| Color | Color Layout |
|-------------------------------|---------------------------------|
| AUTO COLOR CORRELOGRAM [30] | CEDD [12] |
| FUZZY HISTOGRAM [24] | FCTH [13] |
| FUZZY OPPONENT HISTOGRAM [60] | JCD [12] |
| GLOBAL COLOR HISTOGRAM [51] | LUMINANCE LAYOUT [42] |
| OPPONENT HISTOGRAM [60] | MPEG7 COLOR LAYOUT [33] |
| Edge | Structure |
| EDGEHIST [4] | JPEG COEFFICIENT HISTOGRAM [42] |
| MPEG7 EDGE HISTOGRAM [45] | PHOG [9] |
| HOUGH [29] | PROFILE [4] |
| Texture | Other |
| GABOR [42] | BLOCKS [4] |
| HARALICK [25] | COMPACTNESS [45] |
| LOCAL BINARY PATTERN [27] | MAGNOSTICS [4] |
| TAMURA [58] | STATISTICAL NOISE [4] |

Table 1: FDIVE uses 24 feature descriptors. These FDs describe a variety of different image features, such as color, layout, structure and shape [4] allowing for a description of a diverse set of properties.

Feature Space (FS): A feature space describes the set of all feature vectors created by an individual feature descriptor. Additionally, a feature descriptor implies a vector space, called feature space. Thus, each feature descriptor has an associated vector space.

Pattern-based Similarity Measure: We define a *pattern-based similarity measure* as a combination of one feature descriptor and a single distance function (Fig. 4 (left)). The *Similarity Advisor* evaluates the usefulness all possible combinations of an FD and a distance function in their ability to separate the clusters of relevant (\mathcal{L}^+) and irrelevant (\mathcal{L}^-) data items.

In FDIVE, we use a set of norms as distance functions because the SOM learning algorithm requires a similarity measure that can describe a vector space allowing for an adaptation of the cluster prototypes “towards” an input vector. FDIVE uses the following norms: Euclidean L^2 , Manhattan L^1 , $L^{1.25}$ -norm, $L^{1.5}$ -norm and $L^{1.75}$ -norm, which are all L^p -norms with $\|x\|^p = (\sum_{i=1}^d |x_i|^p)^{1/p}$ and the implied metric $d(x, y) = \|x - y\|$ as a similarity measure.

5.1 Comparability of Pattern-based Similarity Measures

Every FD describes a different set of data properties by mapping a data item to a vector representation. To derive useful similarity relations, we need to use a distance function that applies to the vector. We limit ourselves to L^p -norms. However, this approach is extendable to other distance functions and similarity coefficients, including those which do not satisfy the metric axioms.

We leverage the definition of normed vector spaces, which is defined as $(V, \|\cdot\|)$ where V is a vector space and $\|\cdot\|$ a norm on V . We use this definition and apply it to the combination of an FD and its FS along with an L^p -norm with $p \in [1, \infty)$. Throughout this paper, the term distance function refers to the induced metric $d(x, y) = \|x - y\|$. In FDIVE, we define a *pattern-based dissimilarity measure*, a combination of a single FD and a distance function, formally as $dist_d^{FD} : (x, y) \rightarrow [0, \infty)$ with $dist_d^{FD}(x, y) = d(FD(x), FD(y))$ and $x, y \in DS$ data items of the dataset.

We apply a non-standard normalization to transfer a feature space FS and the associated norm into a comparable format. To achieve this outcome, we center the set of all feature vectors $x \in FS$ on the origin, such that the center of each dimension range is located at the origin. This translation does not change vector distances. For this we create a translation vector $t \in \mathbb{R}^n$. The components of t are defined for each dimension i as

$$(1) \quad t_i = 0.5 \cdot (\max_{v \in FS}(v_i) + \min_{v \in FS}(v_i))$$

With this, we can formalize the necessary normalization step to transform the feature space into a comparable state as described by the following function.

$$(2) \quad \text{normalize}(x) = (x - t) / \max_{v \in FS}(\|v - t\|) \text{ with } x \in FS$$

The normalization needs to be performed for all elements x of feature space to convert it into a comparable format. This normalization can be implemented with a complexity of $O(N \cdot M)$ for the full dataset of size N and M *pattern-based similarity measures* implied by the similarity measures, leveraging the mathematical definition of a norm. In essence, this transformation translates all vectors such that the center of each dimension range is located at the origin and scales all vectors such that $\|x\| \in [0, 1]$ for all vectors x , while preserving relative distances between all vectors according to the norm. This normalization allows us to compare the different topologies created by different feature descriptor and norm combinations.

This approach can extend to non-norm similarity coefficients, under the following implications. (1) Ideally, the subsequently applied classification model is compatible with the similarity coefficient, e.g., Self-Organizing Maps require a norm as an internal distance function since prototype vectors need to be updated “towards” an input vector. (2) With non-norm similarity coefficients, the following non-standard normalization needs to be performed. Non-norm similarity coefficients define the difference purely by the distance of data items. This requires the normalization of the full distance matrix of the feature space. This leads to a significant complexity increase since all pair-wise distances need to be computed in $O(N^2 \cdot M)$.

5.2 Quality Metrics for Pattern-based Similarity Measures

In this section, we discuss a set of heuristic quality metrics that we designed to estimate the applicability of a similarity measure for a given analysis task. All quality metrics are calculated based on the transformed features space and the associated distance function, according to the previous section. We measure two concepts, *Inter-Group-Distance*, and *Intra-Group-Distance*. A group is defined as a set data items sharing an identical label, i.e., relevant or irrelevant. Thus one group is formed by all elements in \mathcal{L}^+ and another by \mathcal{L}^- . An intuition is given in Fig. 5.

Inter-Group-Distance measures the similarity of the groups, by calculating synthetic centroids of \mathcal{L}^+ and \mathcal{L}^- , and subsequently determining the distance between both centroids or short $Q_{inter}(\mathcal{L}^+, \mathcal{L}^-) = \text{dist}(\mathcal{L}_c^+, \mathcal{L}_c^-)$. A large *Inter-Group-Distance* is highly desirable.

Intra-Group-Distance measures the maximum distance between distinct elements one of label, i.e. \mathcal{L}^+ and \mathcal{L}^- . Thus, we can say $Q_{intra}(\mathcal{L}) = \max_{i, j \in \mathcal{L}}(\text{dist}(\mathcal{L}_i, \mathcal{L}_j))$, where $i \neq j$. We will apply the above heuristic for every dissimilarity measure.

We experimented with different combinations of *Inter-* and *Intra-Group-Distance* and variants also involving mean and median values instead of the maximum for the *Intra-group-distance*. We also combined both measures into $Q_{comb}(\mathcal{L}^+, \mathcal{L}^-) = Q_{inter}(\mathcal{L}^+, \mathcal{L}^-) - w \cdot (Q_{intra}(\mathcal{L}^+) + Q_{intra}(\mathcal{L}^-))$, with a weighting w . In general, we found that the *Inter-Group-Distance* performed the best on its own, i.e., with $w = 0$.

Other metrics in the context of internal cluster quality metrics use similar notions to *Inter-* and *Intra-Group-Distance*. Cutting et al. [15] describe internal cluster metrics such as the cluster self-similarity defined as the average distance of all cluster members or the average distance of all cluster members to the centroid. We found that this measure did not describe the group separation very well since the ideal case describes a cluster concentrated on a small region. This case rarely occurs in real-world scenarios, without all points of both \mathcal{L}^+ and \mathcal{L}^- clusters being concentrated at the same location. We looked at internal cluster quality metrics such as the Dunn Index [18] which measures the ratio of minimum cluster distance to the maximum cluster extent. Another measure is the Davies-Bouldin index [16] describing the sum of cluster extents to the centroid distances. Both approaches include the notion similar to the *Intra-Group-Distance*. We found that both measures were sensitive to outliers and thus were not as useful as the *Inter-Group-Distance* heuristic.

We use and suggest the *Inter-Group-Distance* on its own in all applications and evaluations of FDIVE. This distance-based score is used to rank the set of similarity in descending order, as shown in Fig. 4 (right). The *Similarity Advisor* shows the score as bar. Additionally, we display the topology of the associated features space. Labeled data items are highlighted, allowing users to verify the separation of relevant and irrelevant data items. With the *Inter-Group-Distance* we found a heuristic that is intuitive, easy to calculate and performs well, as we will show in Sec. 7.2.

6 LEARNING RELEVANCE OF DATA POINTS WITH SELF-ORGANIZING MAPS

FDIVE features a SOM-based classifier, which is used to classify data items by their assignment to a SOM-neuron, and to learn decision planes in the high-dimensional space discriminating \mathcal{L}^+ and \mathcal{L}^- . We introduce a set of visual encodings to guide the user to potentially interesting data subsets, or regions of classifier uncertainty.

6.1 SOM as Visual Classifier

SOMs cluster similar items in cells, which provides users with an intuition about the classification process. SOMs preserve distance relations between cells allowing for orientation in the data space. The tree-structure and SOM cell exploration allow for a drill-down from the data space to clusters and individual data items. SOM cells are arranged in a grid which is directly visualizable, which also applies and our tree-like classifier model. Additionally, our SOM classifier conveys areas of uncertain classifications by highlighting cells with mixed labeling and cells with a low amount of labeled data items. Additional labels improve the classification. Labels can be added in those specific areas. The grid size is a user parameter, and 3×3 is the default setting.

We use Self-Organizing Maps as a basis for our model because it is visually explorable; it partitions the feature space and the data

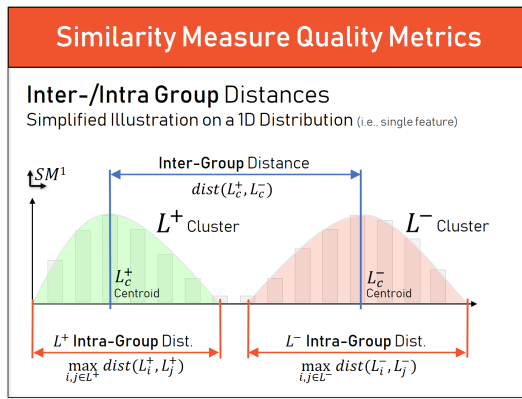


Figure 5: We propose two quality metrics to evaluate similarity measures. *Inter-Group-Distance* describes the distance between the centroids of the relevant and irrelevant data, measuring how well a similarity measure separates both groups. The *Intra-Group-Distance* is defined as the maximum distance in the relevant or irrelevant data, measuring whether a similarity measure describes elements of the same group to be dissimilar.

space, which provides the user with analyzable chunks. The supplied relevance labels and the selected dissimilarity definition are used to calculate a SOM-based relevance model that separates relevant and irrelevant data items. The model can be explored for visual model understanding. Moreover, the model visually conveys areas of uncertainty. The user is then able to refine the relevance feedback in areas of uncertainty, namely the decision boundary. Since our approach is focused on the creation of a relevance model reflecting the user’s notion of relevance and thus in essence, not for exploratory analysis, we limit our approach to the representation of a user’s fixed notion of relevance.

Classifier Training: A regular SOM is likely to create cells in which relevant and irrelevant items are mixed. We resolve this by proposing a hierarchical SOMs that allows for the expression of fine-grained differences in the user’s notion of relevance. For this reason, we merge the concept of a tree with the concept of child SOMs presented by Sacha et al. [53], where a new SOM is calculated only with a subset of the dataset determined by the cell selection of a parent SOM. However, our algorithm creates a classifier automatically without any user interaction other than supplied labels. We automatically calculate a child SOM only from the data items assigned to the given cell c if this cell exhibits a mixture of relevant of irrelevant greater then a threshold m_t , i.e., $MixRatio(c) > m_t$ with

$$(3) \quad MixRatio(c) = \min(|\mathcal{L}_c^+|/|\mathcal{L}_c|, |\mathcal{L}_c^-|/|\mathcal{L}_c|)$$

The cell needs to contain enough data items in order for a child SOM to be useful. We model this circumstance by another threshold value c_t , such that the number of items in a cell $|E_c|$ must exceed c_t . Thus c_t determines the split criterion. In FDIVE, the creation of SOM models is based on the supplied similarity measure, as determined by the *Similarity Advisor*, and the relevance labels. The resulting SOM-based model can exhibit a tree structure (Fig. 6 (1)). We limit the layout to 3×3 to leverage the projection of a SOM into 2D but not handle an excessive amount of children for a given parent in the classification tree.

Classification of Data Items A classification of a given data item is performed recursively, similar to a decision tree. (1) Find the most similar neuron in the root SOM; (2) If the node has a child, perform the same action recursively on the child SOM; (3) If the SOM node has no child, classify the item as the predominant label



Figure 6: Visual Exploration of SOM Model: 1) Classifier tree. 1a) Parent of the currently observed SOM. 1b) Children of the current SOM. 2) Detailed SOM Display. 3) Scatter plot highlighting data of the SOM node.

of the respective cell, i.e., relevant or irrelevant; (4) If no label information is available for this node, use the next most similar cell with label information in that specific SOM.

6.2 SOM Exploration and Refinement

Our SOM-based visual classifier is visually explorable. It conveys its relevance decisions through multiple visual and interactive techniques. The main navigation happens in the visual classifier tree (Fig. 6 (1)). Each SOM can be selected to examine it in detail. The currently active SOM is marked with a purple border. A purple dot highlights the parent SOM of the selected child SOM. The color coding of the grid in each SOM is intuitive, green signals a predominance of relevant items, red a predominance of irrelevant items. Yellow signals a mixture of relevant and irrelevant items, according to the *MixRatio* of a cell. Such cells are likely to be recursively refined, as described in the previous section. We deliberately chose this encoding since it intuitively signals the relevance of data items from green over yellow to red gradient. Fig. 6 (3) shows the classification outcome for data items assigned to a child SOM or individual cell. This allows us to detect whether a cell is on decision boundary.

To provide insight into the data items assigned to each node, we provide a range of stackable cell visualizations that can be selected in a user-defined order.

Relevance Label Quality: The label quality is depicted as colored squares on top of each node. We use the *MixRatio* to determine the color and create a gradient from red over yellow to green; red is representing only irrelevant, green only relevant items within the cell and yellow implies an uncertain cell, i.e., decision boundary. A white dot signalizes that the cell contains not enough labeled data items, visually prompting users for more labels.

Feature Histogram: This layer displays the trained vector of the node. It can be used to judge the differences of SOM cells according to the currently recommended feature description. If the currently active FD is interpretable, like an FD derived from a color histogram, describing the color spectrum of an image, it can also hint at the properties of the contained data items.

The user can also utilize two other layers, the quantization error (QE) [50] and the U-Matrix [59], to explore clusters of nodes that

should be treated similarly by the model. Also, we support the user with detailed information about the number of assigned data items, relevant, irrelevant, and neutral data items. This information allows the user to judge the importance of a given node and the amount of information available to the model.

6.3 Visual Active Learning with SOMs

Cells with a low amount of labeled data items are uncertain. We measure this uncertainty with the *LabelRatio* of a cell c . $|E_c|$ defines the number of items in a cell. Thus, we define the *LabelRatio* as

$$(4) \quad \text{LabelRatio}(c) = (|\mathcal{L}_c^+| + |\mathcal{L}_c^-|) / |E_c|$$

The model marks cells that do not have a child SOM with a white dot if $\text{LabelRatio}(c) < q_t$, where q_t defines a threshold. A white dot signals uncertain neurons with a low label count to prompt the user to supply additional labels in these uncertain data regions. If the user does not supply an additional label by the suggested SOM node, the query formulated by an active learning system is generated from those marked nodes. For every node, the user can request details-on-demand in the form of a model-refinement dialog, similar to the annotation view, presented in Sec. 4.

7 EVALUATION

Approaches involving relevance feedback are not straightforward to evaluate, as the results depend on both hidden and explicit user preferences and the definition of the learning components [49]. Therefore, we show its usefulness by applying it to a real-world use-case. We evaluate the general workflow, including the *Similarity Advisor*, through a comparison to multiple feature selection techniques.

7.1 Case Study: Synapse Detection

The goal of connectomics is to reconstruct the neural wiring diagram from Electron Microscopic (EM) images of the animal brain to improve the understanding of neuropathology and intelligence. A synapse is a functional structure that enables signal transfer from one neuron to the other, which connects individual neurons into a complex network. Manual labeling of synapses can be extremely hard because (1) there are approximately one billion synapses in a 1mm^3 cube of a mouse brain, and (2) the labeling of synapses requires expertise and cannot be crowdsourced. Therefore, a good labeling system of synapses should be semi-automatic and only provide informative samples to the domain experts to improve the labeling efficiency. To showcase the effectiveness of our proposed approach, we applied the annotation system to a high-resolution EM image dataset generated by a multi-beam scanning electron microscope¹. In total, there are 4,000 image patches, half of them containing a synapse at the center of the image, while the other half do not contain synapses. In this study, we show how our system helps experts classify synapse images and non-synapse images without any labeled training set and pre-specified domain knowledge.

CNN-based approaches have achieved state-of-the-art performance on image classification tasks [26, 38]. However, there are still two main shortcomings of CNN-based methods. First, because the model space of CNNs can be huge, the model can easily overfit the training set and have poor performance on the test set, which requires a large training set. Second, the features extracted over convolutional layers are hard to interpret, which restricts the understanding of the discriminative features, especially for scientific applications where the expert wants to have a full understanding of the model.

Thus we perform a case study² involving the classification of Electron Microscopy (EM) images of brain cells. A domain expert is tasked with the creation of a relevance model able to distinguish

images depicting neuronal synapses. The domain expert has experience in the area of connectomics and the interpretation of EM images, including the identification of cell structures such as cell organelles and neuronal synapses. The study was conducted as a semi-structured interview. The case-study was performed after a training period. The expert performed a total of nine iterations to teach our relevance model the difference between EM images containing synapses and those which do not. Fig. 7 shows four key events in the model learning process. After the initial annotation of 40 data items, the system suggested the EDGEHIST FD. The expert finished the first iteration by labeling data items in cells with a white dot. A total of 95 images were annotated as relevant and 65 as irrelevant. In the second iteration, the system suggested the TAMURA FD. The expert labeled 63 images as relevant and 57 as irrelevant. In the third iteration, the system suggested the TAMURA FD again. In the fourth and fifth iteration, the MPEG7 EDGE HISTOGRAM FD was suggested. In iteration six to nine the system consistently suggested the HARALICK FD point at convergence on this specific FD. The expert followed the recommendations of the *Similarity Advisor* in every iteration, finishing after the ninth iteration.

In the first three iterations, the system indicated uncertain cells. In later iterations, we are able to check the distribution of samples in a SOM on the scatter plots to see if they are still mixed up. In the end, it notified the expert that it has enough labels, such that no further inspection or labeling is necessary. After several iterations of labeling, the expert noticed that samples are separated in the classification scatter plot, and, when inspecting the individual nodes pertaining to a data region, the labels of similar data items were matching. From the root node to the leaf nodes, he was able to see a trend towards purity. Therefore, when the uncertainty indicators (i.e., white dot) disappears, the nodes with mixed colors are more appealing to be labeled. The inspection of nodes was helpful to the expert to validate whether a set of samples spread out on the scatter plots and thus do not form a coherent cluster. When inspecting a cell colored in yellow, the expert was able to see decision boundaries. Subsequently, the expert labeled ten queried samples to refine the decision boundary. After labeling one node, the color of the node itself and its sibling nodes may change, and the expert was able to verify the impact. The expert noted that the appearance of the scatter plot changed several times at the initial iteration and that the relevant and irrelevant samples on the scatter plots were mixed and not forming a coherent cluster. However, after several iterations, the model converges to a specific similarity measure, and samples become more separable on the scatter plots.

With FDIVE we can learn to distinguish and extract relevant patterns from a large high dimensional dataset, in this case, EM images depicting synapses, using a sparse amount of labels. Whenever a new label is applied, the system conveys its impact visually. The relevance model is visually explorable and refinable such that the expert was able to assess the model quality and the convergence towards a useful relevance model.

7.2 Quantitative Framework Evaluation

This evaluation compares the best best-breed-competitor generated by 3 algorithms and 4 different FD sizes against our “one-shot” *Similarity Advisor* result. Comparing a recombination of all features with the *Similarity Advisor* using only the predefined feature descriptors make this evaluation biased against our approach. However, we were still able to outperform the best best-breed-competitor in 36 out of 75 cases. We evaluate FDIVE on the following options and parameter settings with the central goal to show the usefulness of ranking pattern-based similarity measures for model learning. We provide a comprehensive overview of the results in Table 2³. The basis for all experiments is the *Quick, Draw!* dataset [32]⁴. We

¹Appendix A provides a visual overview of the Synapse Detection dataset.

²Appendix C shows all intermediate steps in HD images.

³Appendix B contains complete records of all experiments.

⁴Appendix A provides a visual overview of the *Quick, Draw!* dataset [32].

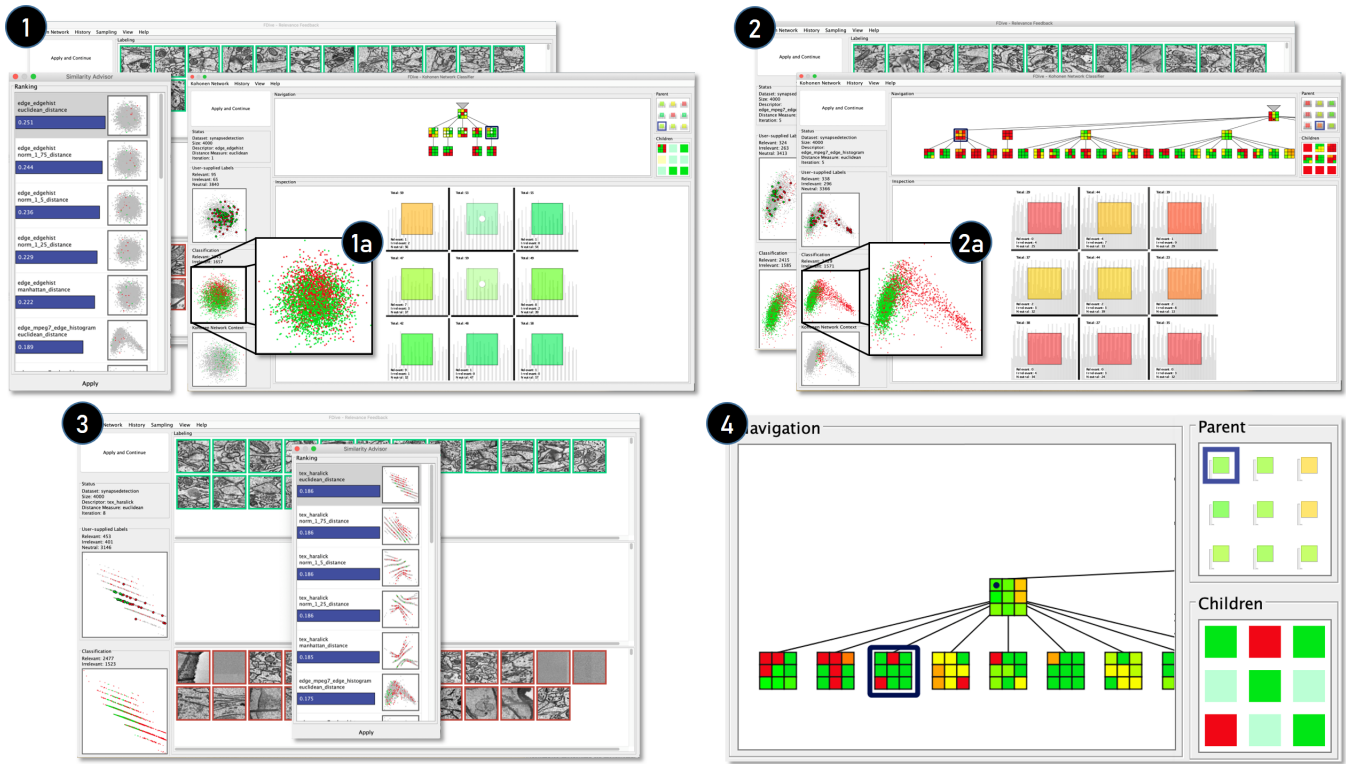


Figure 7: FDIVE learns to differentiate Electron Microscopy (EM) images containing synapses from images that do not. The domain expert found the classification model to be satisfactory after nine iterations. We show four key events in the model learning process. (1) The initial model is classifying the data very poorly, as presented by the scatter plot (1a) being very noisy and mixed. (2) The scatter plot shows a cleaner decision boundary (2a) and the model gets more complex, while the expert labels requested data items. (3) In the seventh iteration, the domain expert noticed that the HARALICK [25] FD combined with the Euclidean distance is recommended for the third time in a row, hinting at convergence for the similarity measure. (4) The last two iterations were spent exploring the model, observing and refining decision boundaries.

reduced the dataset to 4500 images consisting of 150 sketches for each of the 30 labels, describing the depicted objects. We choose the labels *square*, *circle*, *banana*, *crayon*, and *monkey*. These labels cover a variety of shapes with different complexity. We assume each label as a specific analysis target. For each of the target labels, we label progressively more items as relevant and irrelevant. The progression is 25/25, 50/50, 75/75, 100/100, and 125/125 for $\mathcal{L}^+ / \mathcal{L}^-$. This sequence represents an increase in the available labels through the iteration cycle. To verify the validity of the similarity measure ranking, we train a k-NN classification model. We chose k-NN, because it is fully automatic and represents an intuitive classification model. We select three parameters for k , namely 1, 3, and 5. To make our results invariant to the feature selection technique, we conducted our experiments using the ReliefF algorithm, a Linear Ranking Ensemble consisting of ten Recursive Elimination SVMs, and a regular Recursive Elimination SVM. These techniques are described in Sec. 2.1. Those algorithms rank features according to their significance. We choose subsets of different lengths, namely 5, 10, 15, and 20. We perform a feature selection on the concatenation of all FDs (4694 features), resulting in recombination of different features, according to the significance assigned by the feature selection algorithm. This approach creates 12 (= 3 algorithms \times 4 sizes) recombined FDs for each label and label count (i.e., table row). We determine the F_1 score of the trained k-NN for each k with all recombined FDs and all distance functions, yielding 60 (= 12 selected FD \times 5 similarity coefficients) F_1 scores for each k parameter of the k-NN classifier. Table 2 shows the best score out of 60 for a given k in the three columns titled “Best Selected FD”, serving as the benchmark. We compare this score to the single one resulting

from a classification based on the best-ranked similarity measure according to the *Similarity Advisor*. All FDs are in their original state and combined with all available distance functions. The *Similarity Advisor* ranks the similarity measures based on the same label information as available to the feature selection. Table 2 shows the F_1 score for a given k for the best ranked similarity measure in the three rightmost columns titled “Best Ranked Original FD”.

Generally, we found that our the suggested similarity measure performs on a similar level than the best feature selection created by the feature selection algorithms. It outperforms the feature selected FD in all scenarios involving the *banana* label and in 11 out of 15 scenarios pertaining to the *crayon* label. The best-ranked *Similarity Measure* is outperformed in scenarios where the analysis target is a less complex shape (i.e., *square* and *circle*). In case of the *monkey* label, our ranked FD can achieve similar result than the selected FD with 50 or more labeled instance for each of $\mathcal{L}^+ / \mathcal{L}^-$. Given that we compare 60 feature selection-based similarity measures to our single best ranked fixed-FD similarity measure, we can say that the similarity advisor is an efficient and effective method for the evaluation of similarity measures and that the best-ranked measure helps in the creation of a relevance model.

8 DISCUSSION AND FUTURE WORK

With FDIVE, we provide a technique which allows for the iterative learning of a relevance model, including the definition of a useful similarity measure. In the case of FDIVE, a similarity measure comprised of a feature descriptor and a distance function. The visual guidance of the SOM-based relevance model to uncertain classifi-






| Labeling | | Results (larger is better) | | | | | |
|--|--------------------------|----------------------------|---------|---------|-------------------------|---------|---------|
| Target Example | #Labels L^+ / L^- each | Best Selected FD | | | Best Ranked Original FD | | |
| | | Baseline | | | FDIVE | | |
| | | $k = 1$ | $k = 3$ | $k = 5$ | $k = 1$ | $k = 3$ | $k = 5$ |
|  | 25 | .359 | .397 | .410 | .268 | .317 | .312 |
| | 50 | .398 | .464 | .449 | .238 | .330 | .330 |
| | 75 | .326 | .436 | .490 | .215 | .295 | .328 |
| | 100 | .350 | .407 | .465 | .239 | .321 | .347 |
| | 125 | .437 | .516 | .494 | .250 | .328 | .368 |
|  | 25 | .363 | .368 | .320 | .272 | .264 | .239 |
| | 50 | .399 | .444 | .426 | .296 | .292 | .279 |
| | 75 | .461 | .533 | .542 | .286 | .309 | .292 |
| | 100 | .539 | .611 | .567 | .306 | .338 | .323 |
| | 125 | .507 | .600 | .602 | .304 | .357 | .345 |
|  | 25 | .212 | .222 | .224 | .556 | .566 | .490 |
| | 50 | .303 | .310 | .306 | .561 | .574 | .578 |
| | 75 | .323 | .351 | .362 | .605 | .619 | .626 |
| | 100 | .473 | .526 | .507 | .529 | .595 | .586 |
| | 125 | .363 | .447 | .469 | .522 | .585 | .606 |
|  | 25 | .152 | .170 | .187 | .166 | .174 | .153 |
| | 50 | .175 | .157 | .171 | .192 | .216 | .222 |
| | 75 | .180 | .192 | .184 | .197 | .205 | .202 |
| | 100 | .160 | .179 | .186 | .192 | .203 | .194 |
| | 125 | .173 | .186 | .181 | .135 | .142 | .145 |
|  | 25 | .179 | .169 | .173 | .096 | .105 | .101 |
| | 50 | .162 | .165 | .176 | .183 | .247 | .253 |
| | 75 | .197 | .201 | .215 | .180 | .222 | .254 |
| | 100 | .180 | .176 | .191 | .186 | .245 | .273 |
| | 125 | .193 | .209 | .210 | .180 | .235 | .262 |

Table 2: We compare the F_1 scores for different k-NN classifiers. Our heuristic approach performs better for analysis targets with a higher complexity (i.e. *banana*, *crayon* and *monkey*) than state-of-art feature selection algorithms that can draw features from all available feature descriptors (4694 features). It performs worse for less complex patterns (i.e. *square* and *circle*).

cation near decision boundaries improved the understanding and quality of the model. We show that the continuous evaluation of the similarity measure benefits the iterative creation of relevance models, helping them to converge towards increasingly useful results.

One area of improvement noted by the expert was that, upon change of the similarity measure, the relevance model changes its layout, requiring the analyst to relearn it. For this reason, the mapping of different model representations into various feature spaces would allow us to explore the impact of a changed feature space on the model. Making this effect accessible would further the understanding of the feature space and underlying data distribution.

We plan to extend the general concept of the *Similarity Advisor* to other types of distance functions, removing the limitation to vector spaces implied by the L^p Minkowski family of distance measures. This extension would allow us to use other distance functions, such as Cosine, Canberra, or Clark distance. Analysts apply these measures often in specific scenarios and domains. The automatic detection of a distance function would replace the need for an expert, removing the bias introduced through the single fixed distance function. Additionally, we want to explore the application of the

Similarity Advisor in different contexts, such as the validation of feature weightings or the design of feature descriptors based on prototypical representations of the described properties. In this instance, the *Similarity Advisor* could serve as a *concept validator*. Feature descriptors can be linked to visualization types. Through a technique similar to the *Similarity Advisor*, it should be possible to suggest other data representations, such as switching from a scatter plot representation to a parallel coordinate plot. An automatic suggestion of a useful visualization would add another step to a generalized analysis workflow, where many choices an analyst or even system designer can make is automatically assessed and supported. We lay-out the SOM-based relevance model in a tree structure, because it is explainable and an intuitive way of reading a classifier. Techniques introduced by Sacha et al. [53] can be used to enhance its descriptive ability. This addition can lead to novel SOM interactions focused on classification rather than exploratory cluster analysis.

We discuss scalability on two levels. First, we discuss the computational effort of *Similarity Advisor*. The main computational effort lies in the required preprocessing to transform the feature spaces and distance functions into a comparable format. The transformations are parallelizable. The complexity is determined by the dataset size. The complexity of the *Inter-Group-Distance* calculation is determined by the number of supplied labels. However, this relationship is linear. Second, we discuss the scalability limit of the complete FDIVE approach. The main limit approach is the creation of the SOM-based relevance model. However, the results of a previous iteration cycle can be reused in the subsequent cycles. One issue that we found was that the tree representation of the SOM-based model can become very wide. Here we have to consider a tradeoff between the size of the SOM and the associated data partitioning properties and the number of child SOMs leading to a broad tree. We found that a 3×3 SOM is an acceptable size for the SOMs since it is a size where the 2D projection property has a notable effect.

9 CONCLUSION

The extraction of interesting patterns from large high-dimensional datasets is a challenging task. With FDIVE, we present a workflow for the creation of relevance models based on *pattern-based similarity measures*. The system ranks similarity measures according to how well they separate relevant from irrelevant data. Our SOM-based relevance model is interactively exploratory and guides the user to uncertain areas, i.e., decision boundaries. We evaluated our technique with a real-world case study in which we show that FDIVE can reflect the complex differences between electron microscopy images showing synapses of neurons or other brain cell structures. Our comparison to feature selection shows that FDIVE’s *Similarity Advisor* serves as a useful metric to evaluate the discriminative ability of feature descriptor and distance function combinations. With FDIVE, we introduce the concept of continuous *Similarity Advisor* assessment during the learning process of a relevance model. The *Similarity Advisor* concept is applicable to areas where the user expresses his relevance for specific data items and can improve the results of the given task. The full FDIVE approach allows the creation of relevance models for a complex task while providing the user with valuable insights about the learning process, such as the underlying similarity measure and the model properties, including the judgment of classification results in areas of high uncertainty.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the projects A03 of TRR 161 (Project-ID 251654672) and Knowledge Generation in VA (Project-ID 350399414). We thank Michael Blumenschein and the anonymous reviewers for their valuable feedback.

REFERENCES

- [1] M. Ankerst, M. Ester, and H. Kriegel. Towards an effective cooperation of the user and the computer for classification. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 179–188, 2000. doi: 10.1145/347090.347124
- [2] D. Arendt, E. Saldanha, R. Wesslen, S. Volkova, and W. Dou. Towards rapid interactive machine learning: evaluating tradeoffs of classification without representation. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 591–602, 2019. doi: 10.1145/3301275.3302280
- [3] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6:937–965, 2005.
- [4] M. Behrisch, B. Bach, M. Hund, M. Delz, L. von Räden, J. Fekete, and T. Schreck. Magnostics: Image-based search of interesting matrix views for guided network exploration. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):31–40, 2017. doi: 10.1109/TVCG.2016.2598467
- [5] M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck. Feedback-driven interactive exploration of large multidimensional data supported by visual classifier. In *IEEE Conference on Visual Analytics Science and Technology*, pp. 43–52, 2014. doi: 10.1109/VAST.2014.7042480
- [6] J. Bernard, M. Hutter, M. Zeppelzauer, D. W. Fellner, and M. Sedlmair. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):298–308, 2018. doi: 10.1109/TVCG.2017.2744818
- [7] J. Bernard, T. Ruppert, M. Scherer, T. Schreck, and J. Kohlhammer. Guided discovery of interesting relationships between time series clusters and metadata properties. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, 2012. doi: 10.1145/2362456.2362485
- [8] J. Bernard, D. Sessler, A. Bannach, T. May, and J. Kohlhammer. A visual active learning system for the assessment of patient well-being in prostate cancer research. In *Proceedings of the 2015 Workshop on Visual Analytics in Healthcare*, pp. 1–8, 2015. doi: 10.1145/2836034.2836035
- [9] A. Bosch, A. Zisserman, and X. Muñoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pp. 401–408, 2007. doi: 10.1145/1282280.1282340
- [10] S. Bremm, T. von Landesberger, J. Bernard, and T. Schreck. Assisted descriptor selection based on visual comparative data analysis. *Computer Graphics Forum*, 30(3):891–900, 2011. doi: 10.1111/j.1467-8659.2011.01938.x
- [11] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *IEEE Conference on Visual Analytics Science and Technology*, pp. 83–92, 2012. doi: 10.1109/VAST.2012.6400486
- [12] S. A. Chatzichristofis and Y. S. Boutalis. Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In A. Gasteratos, M. Vincze, and J. K. Tsotsos, eds., *Computer Vision Systems*, pp. 312–322. Springer, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-79547-6_30
- [13] S. A. Chatzichristofis and Y. S. Boutalis. FCTH: fuzzy color and texture histogram - A low level feature for accurate image retrieval. In *Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 191–196, 2008. doi: 10.1109/WIAMIS.2008.24
- [14] M. Chegini, J. Bernard, P. Berger, A. Sourin, K. Andrews, and T. Schreck. Interactive labelling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active learning. *Visual Informatics*, 3(1):9–17, 2019. doi: 10.1016/j.visinf.2019.03.002
- [15] D. R. Cutting, J. O. Pedersen, D. R. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 318–329, 1992. doi: 10.1145/133160.133214
- [16] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909
- [17] J. J. Dudley and P. O. Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems*, 8(2), 2018. doi: 10.1145/3185517
- [18] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973. doi: 10.1080/01969727308546046
- [19] E. Eaton, G. Holness, and D. McFarlane. Interactive learning using manifold geometry. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.
- [20] J. Fogarty, D. S. Tan, A. Kapoor, and S. A. J. Winder. Cueflik: interactive concept learning in image search. In *Proceedings of the 2008 Conference on Human Factors in Computing Systems*, pp. 29–38, 2008. doi: 10.1145/1357054.1357061
- [21] M. Gleicher. Explainers: Expert explorations with crafted projections. *Transactions on Visualization and Computer Graphics*, 19(12):2042–2051, 2013. doi: 10.1109/TVCG.2013.157
- [22] R. Gregor, A. Lamprecht, I. Sipiran, T. Schreck, and B. Bustos. Empirical evaluation of dissimilarity measures for 3d object retrieval with application to multi-feature retrieval. In *13th International Workshop on Content-Based Multimedia Indexing*, pp. 1–6, 2015. doi: 10.1109/CBMI.2015.7153629
- [23] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [24] J. Han and K. Ma. Fuzzy color histogram and its use in color image retrieval. *IEEE Transactions on Image Processing*, 11(8):944–952, 2002. doi: 10.1109/TIP.2002.801585
- [25] R. M. Haralick, K. S. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 3(6):610–621, 1973. doi: 10.1109/SMC.1973.4309314
- [26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90
- [27] M. Heikkilä and M. Pietikäinen. A texture-based method for modeling the background and detecting moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):657–662, 2006. doi: 10.1109/TPAMI.2006.68
- [28] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012. doi: 10.1109/TVCG.2012.277
- [29] P. V. C. Hough. Method and means for recognizing complex patterns, December 1962.
- [30] J. Huang, R. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 762–768, 1997. doi: 10.1109/CVPR.1997.609412
- [31] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*. Springer, New York, NY, 2013. doi: 10.1007/978-1-4614-7138-7
- [32] J. Jongejan, H. Rowley, T. Kawashima, J. Kim, and N. Fox-Gieg. Quick, draw!, May 2017. <https://experiments.withgoogle.com/quick-draw>, last accessed 2019-07-15.
- [33] E. Kasutani and A. Yamada. The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *Proceedings 2001 International Conference on Image Processing*, pp. 674–677, 2001. doi: 10.1109/ICIP.2001.959135
- [34] D. A. Keim, J. Kohlhammer, G. P. Ellis, and F. Mansmann. *Mastering the Information Age - Solving Problems with Visual Analytics*. Eurographics Association, 2010.
- [35] B. Kleiner and J. A. Hartigan. Representing points in many dimensions by trees and castles. *Journal of The American Statistical Association*, 76(374):260–269, June 1981. doi: 10.1080/01621459.1981.10477638
- [36] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982. doi: 10.1007/BF00337288
- [37] I. Koprinska. Feature selection for brain-computer interfaces. In *New Frontiers in Applied Data Mining*, pp. 106–117. Springer, Berlin,

- Heidelberg, 2010. doi: 10.1007/978-3-642-14640-4_8
- [38] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.
- [39] J. B. Kruskal and J. M. Landwehr. Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, 37(2):162–168, 1983. doi: 10.1080/00031305.1983.10482733
- [40] X. Lin, F. Yang, L. Zhou, P. Yin, H. Kong, W. Xing, X. Lu, L. Jia, Q. Wang, and G. Xu. A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *Journal of Chromatography B*, 910:149–155, 2012. Chemometrics in Chromatography. doi: 10.1016/j.jchromb.2012.05.020
- [41] Y. Liu and G. Salvendy. Design and evaluation of visualization support to facilitate decision trees classification. *International Journal of Man-Machine Studies*, 65(2):95–110, 2007. doi: 10.1016/j.ijhcs.2006.07.005
- [42] M. Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible java CBIR library. In *Proceedings of the 16th International Conference on Multimedia*, pp. 1085–1088, 2008. doi: 10.1145/1459359.1459577
- [43] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences*, 2:49–55, 1936.
- [44] M. Mehri, R. Chaieb, K. Kalti, P. Héroux, R. Mullot, and N. E. B. Amara. A comparative study of two state-of-the-art feature selection algorithms for texture-based pixel-labeling task of ancient documents. *Journal of Imaging*, 4(8):97, 2018. doi: 10.3390/jimaging4080097
- [45] B. S. Morse. Lecture 9: Shape description (regions). http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MORSE/region-props-and-moments.pdf, last accessed 2019-07-15.
- [46] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre. Clustersculptor: A visual analytics tool for high-dimensional data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pp. 75–82, 2007. doi: 10.1109/VAST.2007.4388999
- [47] D. T. Nhon and L. Wilkinson. Pixsearcher: Searching similar images in large image collections through pixel descriptors. In *Advances in Visual Computing*, pp. 726–735. Springer International Publishing, Cham, 2014. doi: 10.1007/978-3-319-14364-4_70
- [48] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191
- [49] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pp. 109–116, 2004. doi: 10.1145/989863.989880
- [50] G. Pözlbauer. Survey and comparison of quality measures for self-organizing maps. In *Proceedings of the Fifth Workshop on Data Analysis*, pp. 67–82. Elfa Academic Press, 2004.
- [51] Y. Rui, T. S. Huang, and S. Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(1):39–62, 1999. doi: 10.1006/jvci.1999.0413
- [52] T. Ruppert, M. Staab, A. Bannach, H. Lücke-Tieke, J. Bernard, A. Kuijper, and J. Kohlhammer. Visual interactive creation and validation of text clustering workflows to explore document collections. In *Visualization and Data Analysis 2017*, pp. 46–57, 2017. doi: 10.2352/ISSN.2470-1173.2017.1.VDA-388
- [53] D. Sacha, M. Kraus, J. Bernard, M. Behrisch, T. Schreck, Y. Asano, and D. A. Keim. Somflow: Guided exploratory cluster analysis with self-organizing maps and analytic provenance. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):120–130, 2018. doi: 10.1109/TVCG.2017.2744805
- [54] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. P. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, 2014. doi: 10.1109/TVCG.2014.2346481
- [55] Y. Saeys, T. Abeel, and Y. V. de Peer. Robust feature selection using ensemble feature selection techniques. In *Machine Learning and Knowledge Discovery in Databases*, pp. 313–325. Springer, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-87481-2_21
- [56] T. Schreck, D. W. Fellner, and D. A. Keim. Towards automatic feature vector optimization for multimedia applications. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, pp. 1197–1201, 2008. doi: 10.1145/1363686.1363964
- [57] H. Strobel, E. Bertini, J. Braun, O. Deussen, U. Groth, T. U. Mayer, and D. Merhof. Hitsee KNIME: a visualization tool for hit selection and analysis in high-throughput screening experiments for the KNIME platform. *BMC Bioinformatics*, 13(S-8):S4, 2012. doi: 10.1186/1471-2105-13-S8-S4
- [58] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 8(6):460–473, 1978. doi: 10.1109/TSMC.1978.4309999
- [59] A. Ultsch. Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In *Kohonen Maps*. Elsevier, July 1999.
- [60] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010. doi: 10.1109/TPAMI.2009.154
- [61] S. van den Elzen and J. J. van Wijk. Baobabview: Interactive construction and analysis of decision trees. In *IEEE Conference on Visual Analytics Science and Technology*, pp. 151–160, 2011. doi: 10.1109/VAST.2011.6102453
- [62] Z. Wang, Y. Zhang, Z. Chen, H. Yang, Y. Sun, J. Kang, Y. Yang, and X. Liang. Application of relief algorithm to selecting feature sets for classification of high resolution remote sensing image. In *IEEE International Geoscience and Remote Sensing Symposium*, pp. 755–758, 2016. doi: 10.1109/GARSS.2016.7729190
- [63] L. Wilkinson, A. Anand, and R. L. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization*, pp. 157–164, 2005. doi: 10.1109/INFVIS.2005.1532142