

A Visual Active Learning System for the Assessment of Patient Well-Being in Prostate Cancer Research

Jürgen Bernard^{*}
Fraunhofer IGD
Darmstadt, Germany

David Sessler[†]
Fraunhofer IGD
Darmstadt, Germany

Andreas Bannach[‡]
Fraunhofer IGD
Darmstadt, Germany

Thorsten May[§]
Fraunhofer IGD
Darmstadt, Germany

Jörn Kohlhammer[¶]
GRIS, TU Darmstadt
Fraunhofer IGD
Darmstadt, Germany

ABSTRACT

The assessment of patient well-being is highly relevant for the early detection of diseases, for assessing the risks of therapies, or for evaluating therapy outcomes. The knowledge to assess a patient's well-being is actually tacit knowledge and thus, can only be used by the physicians themselves. The rationale of this research approach is to use visual interfaces to capture the mental models of experts and make them available more explicitly. We present a visual active learning system that enables physicians to label the well-being state of patient histories suffering prostate cancer. The labeled instances are iteratively learned in an active learning approach. In addition, the system provides models and visual interfaces for a) estimating the number of patients needed for learning, b) suggesting meaningful learning candidates and c) visual feedback on test candidates. We present the results of two evaluation strategies that prove the validity of the applied model. In a representative real-world use case, we learned the feedback of physicians on a data collection of more than 16.000 prostate cancer histories.

CCS Concepts

•**Human-centered computing** → **Visual analytics**; *Information visualization*; •**Computing methodologies** → *Active learning settings*;

^{*}juergen.bernard@igd.fraunhofer.de

[†]david.sessler@igd.fraunhofer.de

[‡]andreas.bannach@igd.fraunhofer.de

[§]thorsten.may@igd.fraunhofer.de

[¶]joern.kohlhammer@igd.fraunhofer.de

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

VAHC '15, October 25, 2015, Chicago, IL, USA

© 2015 ACM. ISBN 978-1-4503-3671-0/15/10...\$15.00

DOI: <http://dx.doi.org/10.1145/2836034.2836035>

Keywords

Electronic Health Care Data, Visual Analytics, Expert Feedback, Patient Well-Being, Active Learning, Regression Tree, Information Visualization

1. INTRODUCTION

A German proverb says that if you ask two doctors, you get three options. While this can be read as a frustrated remark of a patient, doctors are of course trained to identify all possible diagnoses for the patient's symptoms. However, different expertises of physicians (experience, treatment protocols, education) lead to different estimations of well-being and thus to different suggestions for treatment. For example, family doctors with general knowledge, surgeons and specialists for internal medicine, pathologists, biomedical researchers, might come to different assessments because of their different backgrounds. Furthermore, every course of disease is unique.

This is especially true for the treatment of prostate cancer. Prostate cancer is a highly heterogeneous disease. Many slowly growing tumors never even cause symptoms during the entire lifetime. This requires a careful trade-off of therapeutical effects. Sometimes, however, aggressive cases require immediate intervention. Thus, the distinction between the groups at the earliest possible stage is extremely important. Influencing factors for the progression of prostate cancer may be a mixture of symptoms, predispositions, lifestyle and treatment strategies. The causes for the differences in the individual patients' histories are rarely known completely. Thus, a large amount of Electronic Health Record Data (EHR) is recorded during the disease history. The approach of personalized medicine is to capture the above mentioned differences with the EHR. Another challenge is the individual expertise of physicians treating prostate cancer. A prostate cancer specialist is likely to perform more cases in a week than a physician in a general hospital in an entire year. Thus, the expertise of many physicians is based on the number of cases, which does not necessarily capture the variety of prostate cancer. This is a considerable obstacle for the society to establish or even improve medical best practice. The high-level goal is to offer a leverage point for the improvement of clinical best-practice.

This leads to the question whether the expertise of physicians can be made visible and accessible. In other words, the



Figure 1: The visual active learning system for the assessment of patient well-being. The visual interface at the center contains a patient history visualization (top). A set of sliders at the bottom enables domain experts to enter their assessment of patient well-being for different segments of the history. On the left, the system suggests a set of candidates. Each candidate is visually represented with the patient visualization. On the right, a list of previously learned instances is shown. A color coding (here from green to red) is included in the patient visualizations to represent the well-being scores computed by the active learning model.

tactic knowledge of physicians would become more transparent. Our starting point is to focus on the patient well-being. In essence, the research goal of this work is to create an algorithmic model which learns to assess patient well-being from the feedback of physicians based on an EHR visualization. To create this model, we use an active learning approach. A challenge for the visualization, as well as for the model, is the rich set of attributes potentially relevant to describe the well-being status. In addition, the assessment changes with every new diagnostic event or treatment. Moreover, the interface enabling physicians to give feedback should resemble their everyday practice. Finally, it is a technical challenge to minimize the amount of required assessments and still maximize the generalizability of the trained model.

We present a visual active learning system which learns the well-being status in patient histories. Based on a visualization of a patient’s history, physicians are able to give an assessment in a familiar setting. The manual labels given by physicians are used to iteratively learn a regression tree model for automatic labeling. In every iteration, three coverage models are available to generate a set of unlabeled patient candidates, visually represented in a candidate view. The coverage models implement different criteria to guarantee the suggestion of a diverse set of candidates for active learning. During each iteration, a result view provides direct feedback by showing the scores applied to either labeled or unlabeled data. Users select candidates from either the candidate view for generating new labels, or from the result view for refining previous labels.

The benefit of our system is twofold. On the one hand, it reveals possible inconsistencies in the assessment performed by the same physician on different patients. This is one step toward the long-term goal to offer a training method for physicians. On the other hand, it reveals differences between the assessments of physicians. In this way, the discourse can

be focused on most relevant patient and feature set as exposed by the system.

2. RELATED WORK

Our approach is a method for the iterative improvement of an analytical model for health status assessment based on expert labeling. We structure the related work by specific challenges we needed to solve with a focus on existing approaches from the field of visual analytics. Another important categorization of visual analytics approaches for medicine has been created by Turkay et al. [21]. Turkay et al. distinguish techniques by analytical tasks vs. the level of integration. In their notion, we consider our approach a semi-interactive method for the investigation of dependencies.

Modeling multivariate dependencies The class of dependencies we aim for are often described by regression models. A number of techniques have been presented for visual regression analysis. Guo et al. [13] propose an approach for multi-dimensional linear regression, including a visualization of the parameters in the model space. Piringner et al. [19] present an approach for the visualization and validation of general multidimensional regression models. Both approaches enable the investigation and evaluation of competing model instances. In a recent approach, Mühlbacher and Piringner [18] present an interactive regression modeler. It allows the focused investigation of the model quality and for a local refinement of subspaces, effectively producing a hierarchical, compound regression model. We choose an additive regression tree as the model class because of its flexibility and scalability to the number of independent variables. Interestingly, many recent visualization techniques for multi-dimensional regression are based on multiple two- or three-

dimensional slices. We consider this appropriate for users familiar with mentally reconstructing features of a multidimensional distribution of instances. In contrast, we restrict our approach to the interaction with individual instances (i.e. patient segments) only.

Visual Active Learning A number of concepts have been proposed for adaptive modeling based on user interaction with instances. *Active learning* is such a concept which offers a solution for the classification of sparsely labeled data, and is often used for media classification. With active learning, the machine prompts human experts for information on selected instances. Fogarty et al. [9] use active learning to improve the distance metric used in a system for image retrieval and categorization. While originally a ‘black-box’ approach, visual analytics concepts have been adopted recently to increase the transparency of the modeling process. Seifert and Granitzer’s [20] approach enables the interactive instance selection from a class visualization technique. Hoeflerlin et al. [16] define *inter-active learning* as an extension, which includes the direct manipulation of the classifier and the selection of instances. For a regression manifold model, Eaton et al. [8] propose an interactive editor, with the instances and labels serving as control points. Heimerl et al. [15] propose an active learning approach for training an SVM text classifier. They offer three levels of user interaction: basic active learning, visualization of instances along the classifier boundary, and interactive instance selection. Most visual approaches are using a visualization of the instance distribution along the classifier boundaries respectively regression manifolds. Our approach employs an intermediate strategy between an automatic selection of a single visible instance and a free selection from a distribution of unlabeled data. The user is offered a selection of appropriate candidates by the active learning algorithm. Behrisch et al. [3] present an approach for specifying features of interest in a multiple views of multidimensional data. With the user distinguishing relevant from irrelevant views, the system deduces the preferred views for further exploration.

Visualization of Multiple Patients All components of the approach - instance selection, segment labeling and results - are based on a visualization of EHR on different levels of detail. With the focus on the patient’s history, much EHR visualizations are arranged with respect to a time axis. Lifelines2 [22] is a tool for the visualization and temporal comparison of multiple patient records. It primarily shows events by markers along a timeline. Wilhelm et al. [23] analyze the recovery process of horses (motion capture data) by combining a timeline with projection-based visualization methods. Bade et al. [2] connect time-oriented data and associated knowledge with visualizations at different levels of detail. Analytical systems for entire cohorts often show multiple patients at a higher level of abstraction. Lifeflow [25] and Outflow [24] both support the analysis of sequence patterns in patient histories. Lifeflow focuses on the analysis of frequent patterns using a sequence icicle plot. Outflow targets the backpropagation of events based on the outcome using a directed event graph. Bernard et al. [5] combine time-oriented data with multiple clinical and histological dimensions for the interactive definition of cohorts. Beyond the medical domain, a number of approaches deal with the analysis of time-series segments and sequences. Alsallakh et al. [1] present an approach operating on multivariate time-series segments. In this approach segmentation and labeling are performed by automated techniques, while the ac-

tual model can be selected by the user. Bernard et al. [6] present an approach for the segment and sequence analysis of motion capture data. Motion segments are automatically labeled with a color-coding to represent the similarity between motion sequences. Our approach comprises patient visualizations for different views supporting different tasks. We combine properties of single patient visualizations, to provide details for labeling, and multiple visualizations to offer an overview of the learning progress.

Summary The different components for our concept are closely related to different domains. With our approach we propose a ‘gray-box’ strategy for active learning of a regression model. On the one hand, it offers more context for a domain expert labeling a patient than an active learning approach. On the other hand, it shields the domain expert from the modeling process compared to visual active learning approaches.

3. VISUAL ACTIVE LEARNING SYSTEM

In this section, we present the visual active learning system. We start with a description of the processing workflow. Next, we introduce the visualization of single patients and the implementation of the algorithmic models. Finally, we present the user interface of the active learning system which enables physicians to label the well-being status of patients and to assess the results of the learned models.

3.1 Workflow Description

Raw Data, Segmentation, and Instances The workflow of the visual active learning system is depicted in Figure 2. Patient histories comprise the input of the workflow. These histories consist of multivariate data including time-dependent attributes. A preprocessing step segments the patient histories. The segmentation algorithm uses diagnostic and therapeutic events from the patient history. Each individual segment is treated as an *instance* in the downstream active learning system and consists of a representation of the high-dimensional attribute space that characterizes a patient. Given the potential frequency of events, a segment can be a small time interval. Potentially influential previous events are modeled with additional status variables (e.g., *has been operated*). For temporal progressions within a segment additional attributes are introduced into the feature set. This includes, features like trend, means, and variance. In addition, descriptor-based approaches and different types of data aggregation techniques can be applied to calculate meaningful numerical or symbolical features, for univariate or multivariate time series [17, 7]. In this work, we consider the preprocessing step of the workflow as a black-box approach. However, a meaningful segmentation is key for the usefulness of downstream analysis models [12]. This is why we designed the segmentation algorithm with careful consideration of suggestions expressed by the domain experts. The preprocessing step may include further tasks such as data cleansing, normalization, or feature extraction [11, 4]. The result of the preprocessing phase is a collection of patient histories enriched with a set of segments (instances).

Visual Active Learning The next step of the workflow is the visual active learning system. A visual interface allows domain experts to label the segments, e.g., with the patient well-being status. For this purpose a visual representation of a patient history is provided together with interactive controls that enable manual labeling of the segments. As a

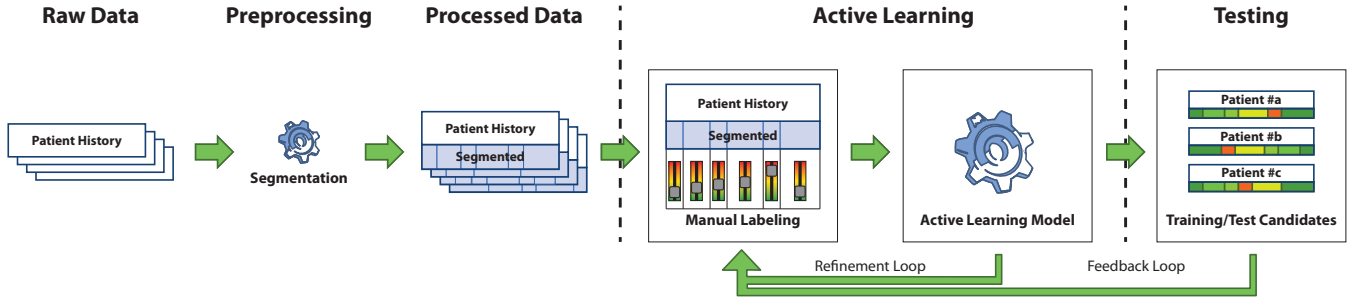


Figure 2: Schematic illustration of the proposed workflow.

next step the manual labels are used as training data for the Active Learning Model (ALM). As a result, the ALM is able to compute a label for every instance in the data set. The accuracy of the labeling depends on the chosen learning model, the manual label information given by the domain expert, the selection of meaningful learning candidates, and the available attributes inside the segmented data. In this iterative process, domain experts are able to assign additional labels to a variety of patients. We refer to this iterative process as the *refinement loop* (cf. Figure 2).

Testing The workflow is concluded with a step for testing the active learning results. For each patient and each segment the label information can be analyzed. The feedback gained from analyzing the results can be used to further enhance the active learning process by additional manual labels. We call this process the *feedback loop* (cf. Figure 2).

3.2 Data Representation and Preprocessing

Patient Histories The system accesses a database with about 16,000 histories of prostate cancer patients. One of the most fundamental requirements to the system was enabling physicians to provide meaningful manual label information for the candidates, i.e., the segments of patient histories. This is why we included a patient visualization which was already in use in the physicians’ current EHR system [5]. We adapted this glyph-based visualization with respect to the data and tasks addressed in this work. For this purpose, we included data attributes relevant for the active learning approach as recommended by the domain experts. The result is illustrated in Figure 3. Cancer patients typically pass through different phases influenced by biological indicators and treatment events. Three of the most relevant attributes stemming from biopsy after operation are visualized at the top in white text boxes (Gleason score, Pt status, Pn status). The most relevant biological indicator is the progression of PSA (a prostate-specific antigen enzyme) which we represent with a black linechart. Another important biological indicator is the occurrence of metastases, encoded with a red-tinted background color in the linechart. We visualize the treatment history with colored horizontal bars at the center of the visualization. Most relevant treatments are operation (pink), hormone therapy (lilac), radiation therapy (cyan), and chemotherapy (purple). In addition, we visualize the number of treatments with a gray staircase metaphor at the bottom of the patient history. Finally, we indicate individual segments with vertical lines and small dark triangles at the bottom.

Segmentation of Patient Histories A patient history

consists of many attributes. Most of these attributes are highly relevant for the assessment of patient well-being. In addition the domain experts expect that the data contains latent indicators, still to be investigated in future. Furthermore, it can be assumed that combinations of observations have an influence on well-being. The temporal development of many attributes contributes to the challenge to assess patient well-being. The segmentation routine proposed in the conceptual workflow eases challenges caused by the temporal domain (cf. Figure 2). We apply the segmentation routine on any patient history of the database. In consequence, we receive *segments* (instances) which represent small temporal intervals within the patient history. We consider the state of a patient history constant within each segment. Thus, every notable change in a patient history triggers the segmentation algorithm to produce a new segment. Each segment characterizes the patient state with a high-dimensional set of features consisting of numerical, ordinal and categorical variables. The set of features is not limited to the attributes described earlier. In fact, new features can be added to the feature set to improve the quality of the ALM. The segments (instances) will be used by the downstream models of the active learning system (cf. Figure 2).

3.3 Model Training

In the following, we describe the ALM of the visual active learning system in detail. In addition, we outline the functionality of our Coverage Models which provide sets of meaningful training candidates.

Active Learning Model The overall goal of the ALM is to compute a label for any instance in the set of given patients. The ALM accepts manual label information as input, e.g., provided by users who label instances with a visual-interactive control element. A review of the related work reveals the variety of different models available for different types of learning approaches (cf. Section 2). The label information of the system is a quantitative variable representing the well-being status of a segment of a given patient history. The underlying feature set of every segment is a mixed data collection consisting of numerical, ordinal, and categorical attributes. Two strategies for coping with mixed data are a) to convert all attributes in a unified format, or b) to apply different models and combine their output by incorporating a weighting concept. We choose strategy a) and bin numerical attributes to categories. According to the required numerical output, we choose a regression tree as a fast decision tree learner. We choose the REPTree implementation in the WEKA data mining software [14]. The model uses the information gain and variance reduction as the splitting criterion. Regression tree logic is used for the creation

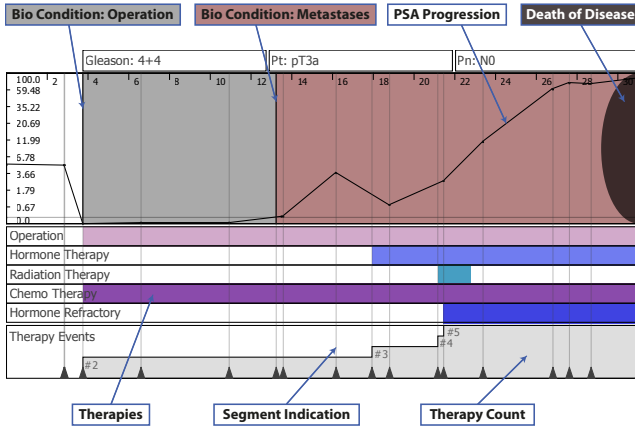


Figure 3: Visual representation of a patient history. We extended the patient visualization of the physicians’ current EHR system [5] with additional attributes relevant for the active learning task. At the top, three static biopsy attributes are shown. Just below the temporal biological attributes are depicted including the PSA value progression. In addition, the event histories of five important therapies are visualized. At the bottom the number of therapies is shown. Finally, the segmentation of the history is indicated with gray vertical lines.

of multiple trees in different iterations in order to seek for the best candidate tree. The pruning strategy of the tree is based on reduced-error pruning with backfitting. Missing values are dealt by splitting the segments into pieces. The REPTree is fast, robust, and suitable for mixed input and a numeric output data. We refer to Section 4 where we compare cross-validation results of the REPTree implementation with implementations of Decision Stump, LinearRegression, M5P, and ZeroR.

Coverage Model A typical challenge of active learning is to identify the next instance to be labeled. Similarly, questions arise when a training process is sufficiently advanced to be finished. The rationale of our Coverage Models is the selection of instances which offer the most information for improving the ALM. Thus, the learning process becomes more efficient and effective. To that aim, the models estimate the ratio of information already used for training. Coverage Models do not relieve the user entirely from the manual selection of examples. In fact, the models provide a set of most promising candidates, which can be further refined by the user as examples for manual labeling. In their survey on this topic, Fu et al. [10] distinguish approaches by the trade-off between eliminating uncertainty and instance diversity. In particular, we aim at choosing the instances for manual labeling in a way that the diversity of labeled data is maximized. This criterion can be defined for single observations C_1 , or for tuples of occurring observations between different attributes C_2 . In addition, we consider that instances with frequently occurring observations are preferred and applied several times, if necessary C_3 . The variance of instances can be characterized with different criteria. The three implementations of the Coverage Model are as follows.

- *Attribute Bin Single Coverage* (C_1)
- *Attribute Bin-Tuple Single Coverage* (C_1, C_2)
- *Attribute Bin-Tuple Frequency Coverage* (C_1, C_2, C_3)

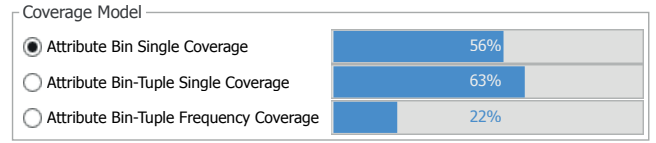


Figure 4: Three different Coverage Models provided with the system. The models estimate the ratio of information gained from the manually labeled data. For example, the first model computed that 56% of the observations in the data set have been learned. Just as well, the Coverage Models select a set of unused candidates which will offer the most information for improving the ALM.

A user is able to change the Coverage Model at any time (see Figure 4). To aid the selection, we estimate the ratio of labeled instances with respect to the criteria of each Coverage Model. Thus, a user may balance the selection based on different criteria. With the coverage scores of the three Coverage Models, we are also able to identify instances which would increase the coverage of the models. By providing a set of appropriate candidates, we tackle the challenge of identifying the next instances to be labeled. The expert is able to select a candidate for manual labeling.

3.4 User Interface

Figure 1 provides an overview of the visual-interactive system. In the following, we describe the three main views of the system in detail, i.e., the Candidate View, the Segment Labeling View, and the Result View.

Candidate View On the left the Candidate View provides a set of candidates carrying information which was not learned before. Every candidate is represented with the adapted patient visualization already used by the physicians. This enables the domain experts to identify relevant disease patterns in individual patient histories without effort. The selection of meaningful candidates is provided by the Coverage Models described earlier. A greedy algorithm defines the set of patient candidates with respect to the instances propagated by the Coverage Model. The user is able to steer the set of provided candidates by choosing a Coverage Model in an interface at the top (also see Figure 4). A click at one of the preferred patients shifts the patient into the Segment Labeling View.

Segment Labeling View At the center, the Segment Labeling View enables physicians to assign well-being labels to segments of a patient history. For this purpose, we apply a large-scale variant of the adapted patient visualization. Slider controls at the bottom of the view enable physicians to assign label information for individual segments. Based on the inquiry of the physicians, well-being labels are assigned in an absolute value domain. Low label values indicate healthy well-being while high values mark life-threatening situations. With the absolute value domain the manual labels of different patients become comparable. We observed physicians assigning the labels very accurately. In a consultation the physicians stated that the various amounts of disease patterns require fine-grained labeling. We support the labeling process with a slider scale from 0 to 10 to give a broad indication (see Figure 1). The thumb of each slider changes its color based on the chosen colormap and the chosen well-being label. New label information triggers the refinement loop of the ALM.

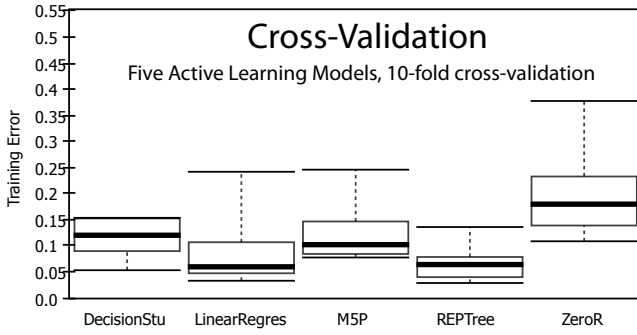


Figure 5: Cross-validation of five learning models. The average fit of each model is measured for a given training set. The REPTree performs well with respect to both a low mean error and a low variance.

Result View The Result View on the right of the system shows a number of candidates with the learned label information. The list-based interface extends patient visualizations with a colored bar representing the well-being scores calculated by the ALM. After each new learning iteration, we recalculate the ALM and update the colors in the patient histories, accordingly. This enables users to follow the learning process. With the feedback collected through the Result View, physicians are able to assess the model quality. The experts may want to re-label the shown patient histories. The selection of a patient history shifts the patient back to the Segment Labeling View, which triggers the feedback loop. At the upper right, we provide a colormap control to enable users to work with their preferred colormap reflecting their notion of well-being. By exposing this parameter to the user, we also provide a means to overcome issues caused by color-blindness.

4. EVALUATION

We evaluated the generalizability of the ALM, i.e., the regression tree. Our goal was to estimate the accuracy of the regression tree implementation in comparison with other models. In addition, we wanted to assess the performance of the winning model for different sets of user labels. In both cases, we applied cross-validation. In our approach, a given set of labeled segments is partitioned into ten disjunct subsets (10-fold cross-validation). In ten iterations, one random segment i is selected for testing while the remaining segments are used to train the model. By doing so, we approximate the exhaustive ‘leave-p-out cross-validation’ strategy. The result of every iteration is the mean difference between the test labels and the output of the regression tree. Since the value domain of the well-being score is quantitative, we can express the quality of the model with a numeric variable.

Cross-Validation of Different Models The first approach is cross-validation applied on different learning models to evaluate the choice of the model. Figure 5 shows the cross-validation results of five different learning models for a given training set. As a result of the performance comparison, we choose the REPTree as the default learning model for the system since the quality results of the REPTree implementation indicate both a low mean error and a low variance. To ensure the statistical significance of our decision with respect to the sample size, we apply a statistical homogeneity test for quantitative distributions (pairwise tests, independent distributions). The Mann-Whitney U Test assesses whether two

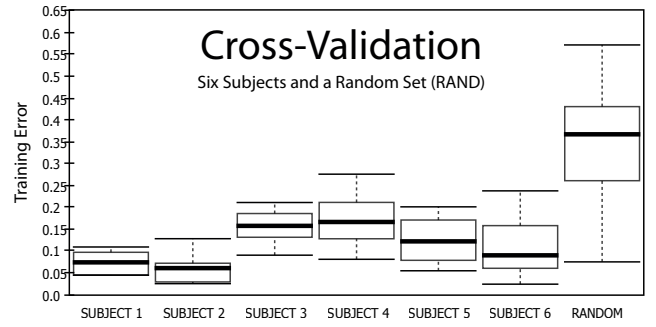


Figure 6: Cross-validation of the REPTree. The average fit is measured for seven training sets. A set of random labels scores worse. Three subjects were asked to assign manual labels in two sessions each, yielding different cross-validation results.

independent samples have equal means. In this respect, the quality results of REPTree outperform three of four competitors significantly ($p\text{-value} \leq 0.005$). the LinearRegression model may be a meaningful alternative ($p\text{-value} = 0.6501$). However, it poses a higher variance compared to REPTree.

Cross-Validation of Different Sets of Labels The second strategy is cross-validation of different sets of labels applied on the preferred model. In Figure 6, we compare the cross-validation results of the REPTree for seven different test sets. Six of the test sets are sets manually labeled by users. As a null hypothesis, a random set (seventh set) is supposed to perform significantly worse. In fact, the Mann-Whitney U Test assesses statistical homogeneity between the random set and the six sets of subjects ($p\text{-value} \leq 0.004$). Interestingly, the six manually labeled sets yield different cross-validation results. It will be an interesting aspect of future work to carry out studies about the dependencies of different mental models (training sets), features sets, and the learning model quality.

5. CASE STUDY

We conducted a case study with physicians of an internationally renowned prostate clinic. After a small introduction to the system, participants were requested to set at least 50 manual labels. To provide generalizability, the physicians were instructed to query at least ten different patient candidates. Our overall goal was to challenge the feature set extracted for the patient segments to validate the usefulness of the approach. We show our findings through the example of a single patient history logged in the course of a single learning session. The patient history shows various interesting biological and therapy events (see Figure 7). The first remarkable event was the *operation* early in the history. As a result of the operation the PSA value dropped almost to zero indicating that the operation was to some degree successful. However, the pathologic result after the operation showed a bad Gleason score (4+5). In the following, the PSA value increased until a *hormone therapy* was conducted. The therapy was successful, again leading to a PSA decrease down to almost zero. However, a year later though, a *relapse* of the PSA value (a warning signal) evolved into a strong *PSA increase*. At the end of the history, the patient is unfortunately in a negative health state. In Figure 7 the patient history is shown together with the colored output of the ALM. The colormap chosen by the participant spans from green (pa-

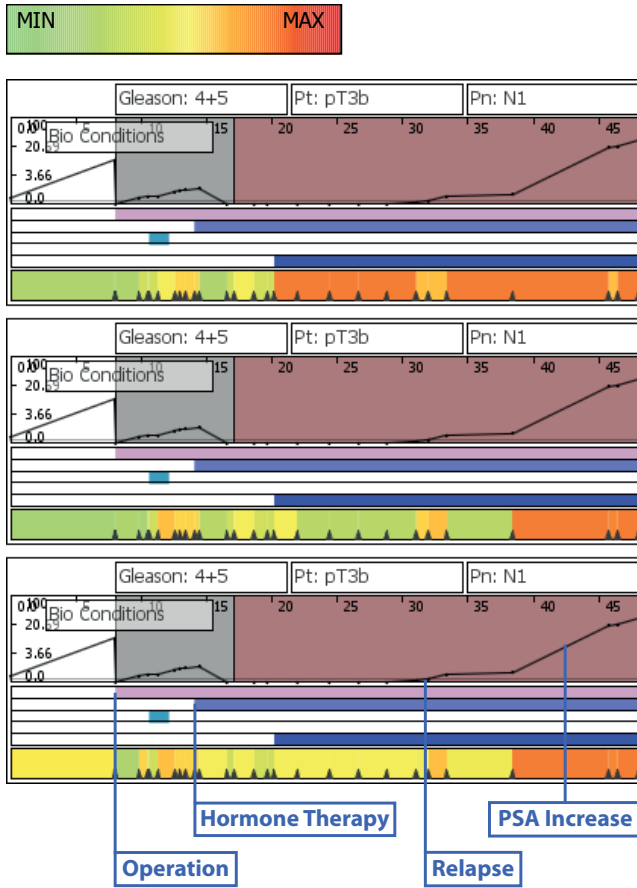


Figure 7: Observation of the model output after the 1st, 5th and 9th learning iteration. The colored bars at the bottom of the three patient visualizations indicate the well-being score learned by the active learning system. In the course of the session the active learning model adapted to the mental model of the physician. The physician affirmed that the final learning result largely matches her assessment.

tient in good condition) to red (bad well-being). The three different visualizations show the learning process after one, five, and nine learning iterations.

Early learning state It can be seen in the top image that the first segment of the patient is labeled green, in spite of the presence of a high PSA value. The physician remarked that the well-being score after the successful hormone therapy (in the middle of the history) was too bad (dark orange) in comparison with the patient well-being condition assessed by the expert. In addition, the increase of the PSA (an indicator for a more severe health condition) led to a better well-being score (light orange) which is contradicting the mental model of the physician. To summarize, the first image shows an inaccurate state of the ALM.

Medium learning state In the image at the center, the calculated patient well-being correlates with the absolute PSA value and, interestingly, with the PSA trend. This insight led the physician to the statement that the system “acts in the way of my thinking”. The finding that the trend of the PSA progression partially beats the absolute value of the PSA is to some degree surprising. We did not expect that an increase of this biological indicator has more weight

than a constant progression, but this explains the manual feedback provided by the physician.

Final learning state We can see at the bottom of the image that the well-being scores of most segments of the history worsened. The answer of the physician to this finding was obvious. After having trained with more patients, the model became more robust and generalizable, indicating this patient as one candidate with a rather unfortunate course of disease. The end of the patient history clearly shows a life-threatening well-being score.

We summarize the results gathered from the observation of all participants of the case study. While the scores of the ALM fits quite well for many patients, the model still has potential for improvement. This becomes particularly obvious for patients with special or unique disease patterns. In addition, we were able to identify specific situations within patient histories which were difficult to cover with the current set of features. Examples are a) the separation of the PSA progression before and after the operation, b) a better characterization of therapy success, or c) the integration of medication information. In these situations the active learning results deviated from the mental model of the physicians. This identifies courses of disease which are currently insufficiently covered by the feature set derived from the EHR. In essence, a more detailed representation of therapeutic events and the involvement of additional features will further improve the model output.

6. CONCLUSION

We presented a visual active learning system which enables physicians to rate the well-being status of patient histories. An active learning model computes labels of well-being scores for all instances of the data collection. The system includes a visualization that enables physicians to receive direct feedback from the current model output. In addition, physicians can choose between different coverage models for estimating the number of instances needed to complete the learning process. To further enhance the learning process the system provides a visual-interactive interface which propagates a set of the most appropriate instances as a suggestion for manual learning. We presented different evaluation strategies for the choice of the model and for assessing the model quality for different subjects and sets of labels. In a case study, physicians were able to learn models which matched their mental models in large parts. In addition, we were able to identify new features which can be included for future improvement.

Future work will include the improvement of the feature set. Furthermore, new classes of models such as Gaussian processes can be introduced to the approach. One of the next research goals will be to enhance the system as a means to compare the mental models of different physicians. Finally, it will be worthwhile to seek for relations between different models, users, and features sets. We assume that this additional design space yields different solutions for different analytical tasks. It will be interesting to apply the tool to facilitate data-driven research in future collaborations.

7. ACKNOWLEDGMENTS

We thank Thorsten Schlomm, Burkhard Beyer, and Dirk Pehrke from the Martini-Clinic Prostate Cancer Center at the University Medical Center Hamburg-Eppendorf (UKE) for their valuable feedback for both the data characterization and the conducted case study.

8. REFERENCES

- [1] B. Alsallakh, M. Bögl, T. Gschwandtner, S. Miksch, B. Esmail, A. Arnaout, G. Thonhauser, and P. Zöllner. A visual analytics approach to segmenting and labeling multivariate time series data. In M. Pohl and J. C. Roberts, editors, *EuroVis Workshop on Visual Analytics (EuroVA)*, pages 31–35. Eurographics, Eurographics, 2014.
- [2] R. Bade, S. Schlechtweg, and S. Miksch. Connecting time-oriented data and information to a coherent interactive visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 105–112, New York, NY, USA, 2004. ACM.
- [3] M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck. Feedback-driven interactive exploration of large multidimensional data supported by visual classifier. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 43–52, 2014.
- [4] J. Bernard, T. Ruppert, O. Goroll, T. May, and J. Kohlhammer. Visual-interactive preprocessing of time series data. In A. Kerren and S. Seipel, editors, *SIGRAD*, volume 81 of *Linköping Electronic Conference Proceedings*, pages 39–48. Linköping University Electronic Press, 2012.
- [5] J. Bernard, D. Sessler, T. May, T. Schlomm, D. Pehrke, and J. Kohlhammer. A visual-interactive system for prostate cancer cohort analysis. *Computer Graphics and Applications, IEEE*, 35(3):44–55, 2015.
- [6] J. Bernard, N. Wilhelm, B. Krüger, T. May, T. Schreck, and J. Kohlhammer. MotionExplorer: Exploratory Search in Human Motion Capture Data Based on Hierarchical Aggregation. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2257–2266, 2013.
- [7] J. Bernard, N. Wilhelm, M. Scherer, T. May, and T. Schreck. Timeseriespaths: Projection-based explorative analysis of multivariate time series data. *Journal of WSCG*, 20(2):97–106, 2012.
- [8] E. Eaton, G. Holness, and D. McFarlane. Interactive learning using manifold geometry. In *AAAI Fall Symposium: Manifold Learning and Its Applications*, volume FS-09-04. AAAI, 2009.
- [9] J. Fogarty, D. Tan, A. Kapoor, and S. Winder. Cueflik: Interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 29–38, New York, NY, USA, 2008. ACM.
- [10] Y. Fu, X. Zhu, and B. Li. A survey on instance selection for active learning. *Knowledge and Information Systems*, 35(2):249–283, 2013.
- [11] T. Gschwandtner, J. Gärtner, W. Aigner, and S. Miksch. A taxonomy of dirty time-oriented data. In G. Quirchmayr, J. Basl, I. You, L. Xu, and E. Weippl, editors, *CD-ARES*, volume 7465 of *Lecture Notes in Computer Science*, pages 58–72. Springer, 2012.
- [12] T. Gschwandtner, H. Schuman, J. Bernard, T. May, M. Bögl, S. Miksch, J. Kohlhammer, M. Röhlig, and B. Alsallakh. Enhancing time series segmentation and labeling through the knowledge generation model. In R. Maciejewski and F. Marton, editors, *Poster Proceedings of the Eurographics Conference on Visualization (EuroVis 2015)*, page 3. The Eurographics Association, 2015.
- [13] Z. Guo, M. Ward, and E. Rundensteiner. Model space visualization for multivariate linear trend discovery. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 75–82, 2009.
- [14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.
- [15] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2839–2848, 2012.
- [16] B. Höferlin, R. Netzel, M. Höferlin, D. Weiskopf, and G. Heidemann. Inter-active learning of ad-hoc classifiers for video visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST), 2012*, pages 23–32, 2012.
- [17] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, DMKD '03, pages 2–11, New York, NY, USA, 2003. ACM.
- [18] T. Mühlbacher and H. Piringer. A partition-based framework for building and validating regression models. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):1962–1971, 2013.
- [19] H. Piringer, W. Berger, and J. Krasser. HyperMoVal: Interactive Visual Validation of Regression Models for Real-Time Simulation. *Computer Graphics Forum*, 2010.
- [20] C. Seifert and M. Granitzer. User-based active learning. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 418–425, 2010.
- [21] C. Turkay, F. Jeanquartier, A. Holzinger, and H. Hauser. On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In A. Holzinger and I. Jurisica, editors, *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, volume 8401 of *Lecture Notes in Computer Science*, pages 117–140. Springer Berlin Heidelberg, 2014.
- [22] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith. Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1049–1056, 2009.
- [23] N. Wilhelm, A. Vögele, R. Zsoldos, T. Licka, B. Krüger, and J. Bernard. Furyexplorer: visual-interactive exploration of horse motion capture data. In *Proceedings of SPIE, Visualization and Data Analysis (VDA)*, pages 93970F–93970F–15, 2015.
- [24] K. Wongsuphasawat and D. Gotz. Outflow: Visualizing patient flow by symptoms and outcome. In *IEEE VisWeek Workshop on Visual Analytics in Healthcare, Providence, Rhode Island, USA*, 2011.
- [25] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman. Lifeflow: Visualizing an overview of event sequences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1747–1756, New York, NY, USA, 2011. ACM.