



Interactive visual labelling versus active learning: an experimental comparison

Mohammad CHEGINI^{†1,2}, Jürgen BERNARD^{†3}, Jian CUI², Fatemeh CHEGINI⁴,
 Alexei SOURIN^{†2}, Keith ANDREWS^{†5}, Tobias SCHRECK¹

¹*Institute of Computer Graphics and Knowledge Visualisation, Graz University of Technology, Graz 8010, Austria*

²*School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore*

³*InfoVis Group, University of British Columbia, Vancouver V6T1Z4, Canada*

⁴*Max Planck Institute for Meteorology, Hamburg 20146, Germany*

⁵*Institute of Interactive Systems and Data Science, Graz University of Technology, Graz 8010, Austria*

[†]E-mail: m.chegini@cgv.tugraz.at; jubernar@cs.ubc.ca; assourin@ntu.edu.sg; kandrews@tugraz.at

Received Oct. 6, 2019; Revision accepted Jan. 17, 2020; Crosschecked Jan. 30, 2020

Abstract: Methods from supervised machine learning allow the classification of new data automatically and are tremendously helpful for data analysis. The quality of supervised machine learning depends not only on the type of algorithm used, but also on the quality of the labelled dataset used to train the classifier. Labelling instances in a training dataset is often done manually relying on selections and annotations by expert analysts, and is often a tedious and time-consuming process. Active learning algorithms can automatically determine a subset of data instances for which labels would provide useful input to the learning process. Interactive visual labelling techniques are a promising alternative, providing effective visual overviews from which an analyst can simultaneously explore data records and select items to a label. By putting the analyst in the loop, higher accuracy can be achieved in the resulting classifier. While initial results of interactive visual labelling techniques are promising in the sense that user labelling can improve supervised learning, many aspects of these techniques are still largely unexplored. This paper presents a study conducted using the mVis tool to compare three interactive visualisations, similarity map, scatterplot matrix (SPLOM), and parallel coordinates, with each other and with active learning for the purpose of labelling a multivariate dataset. The results show that all three interactive visual labelling techniques surpass active learning algorithms in terms of classifier accuracy, and that users subjectively prefer the similarity map over SPLOM and parallel coordinates for labelling. Users also employ different labelling strategies depending on the visualisation used.

Key words: Interactive visual labelling; Active learning; Visual analytics

<https://doi.org/10.1631/FITEE.1900549>

CLC number: TP311

1 Introduction

Labelling is assigning a class from the label alphabet to an instance (a record) in a multivariate dataset. Supervised machine learning algorithms, such as classifiers (Bishop, 2006), must be trained on

a labelled dataset to perform. These methods learn how to generalise new data based on existing known data examples, which are provided with a class label. Creating a training dataset is essential to find a small subset of a dataset that delivers the best accuracy for the classifier. Although labelling a dataset is necessary, it can be a dull, time-consuming, and expensive task.

[‡] Corresponding author

ORCID: Mohammad CHEGINI, <https://orcid.org/0000-0002-3516-8685>

© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2020

To address this problem, active learning can

help the analyst by suggesting instances to label (Settles, 2009). Active learning algorithms effectively reduce the number of records that need to be interactively labelled. Active learning techniques require heuristics for record selection, which often depend on the classification problems or data characteristics. Furthermore, the interactive visual labelling (VIAL) (Bernard et al., 2018c) tools build explorable visual overviews on top of active learning algorithms and can outperform classic active learning techniques in term of accuracy (Bernard et al., 2018a). Such combined tools allow an analyst to label a multivariate dataset in a visual environment, while receiving feedback and guidance from the system. Based on the overall data characteristics perceived by the analyst, conscious choices can be made as to what distinguishes groups of data, how many groups there should be, and how many representative records can be labelled. Immediate feedback can be given regarding the current set of labelled records, for example, by visualising changes and improvements to the given classifier in response to the given changes in labelling. Thereby, users can also obtain an understanding of which choices affect the classifiers, and hence contribute to understandable and explainable machine learning models.

Since there are multiple visualisation and interaction techniques, the following research question arises: how do characteristics of these techniques and datasets affect performance and user experience for interactive visual labelling tasks? This key question will be broken down into several sub-questions in Section 4.1. To address them, this paper describes a comparative user study of three well-known interactive visualisation techniques for visual labelling: similarity map, scatterplot matrix (SPLOM), and parallel coordinates (Inselberg, 1985). Using the existing mVis visual data exploration tool (Chegini et al., 2019a), nine machine learning experts labelled two multivariate datasets in each of these three views separately. The quantitative measures from these tasks are accumulated and compared to each other and to active learning algorithms. In addition, the techniques are compared to each other in terms of user experience. The results confirm that involving the user in labelling using visual exploration facilities can improve the machine learning process and enhance model understanding.

2 Related work

Semi-supervised machine learning algorithms require some initial labelled instances (data records), and later the system acquires further labelled instances with the help of the oracle (i.e., analyst). Active learning (AL) strategies provide guidance by asking the analyst to label those instances, which might provide better differentiation. Common active learning strategies include smallest margin (Scheffer et al., 2001; Wu et al., 2006), entropy-based sampling (Settles and Craven, 2008), and least significant confidence (Culotta and McCallum, 2005). These three strategies are fast, and are commonly used as uncertainly sampling active learning strategies (Bernard et al., 2018b). For the robustness of the experiment, in this paper, all three techniques are included in the comparison with interactive visual techniques.

In contrast, supervised machine learning algorithms require a sufficient number of labelled instances at the beginning. Classification techniques, such as random forest (Ho, 1995), are among these algorithms. Classifiers are an essential part of both active learning algorithms and interactive visual labelling strategies. Classifiers are used to provide visual feedback to the user during interactive labelling. In this paper, to remove any potential bias, random forest is used as the classification technique for both active learning algorithms and interactive visual labelling techniques.

To date, multiple strategies for interactive visual labelling have been proposed. For example, Heimerl et al. (2012) incorporated active learning for interactive visual labelling of text documents. Höferlin et al. (2012) introduced interactive learning, which extends active learning to a visual analytics process for building ad hoc training classifiers. Bernard et al. (2018c) proposed VIAL, a unified process combining model based AL strategies with visual analytics techniques. Interactive visual labelling strategies integrate various machine learning and visual analytics strategies to label an unlabelled dataset, so it can be used to train a machine learning model. Bernard et al. (2018a) ran an experiment to show that interactive visual labelling strategies can outperform pure active learning algorithms in terms of performance and accuracy. Later, Chegini et al. (2019a) integrated interactive labelling into mVis, a tool built

based on previous work by Shao et al. (2017) and Chegini et al. (2018). mVis provides visual analysis of high-dimensional data using multiple coordinated views, including similarity maps, SPLOM, and parallel coordinates. Interactive labelling functionality of mVis allows users to create and name groups (classes) and add instances to them. Filtering and colour-coding support efficient comparison of the labelled groups across the different views. In a preliminary study (Chegini et al., 2019b), mVis was found to be intuitive and usable, helping analysts gain insight into their data, and hence provided a sound technical basis for the comparative study.

3 Methods

Using mVis, the performance of three different visualisation techniques for labelling a multivariate dataset was compared. Fig. 1 shows mVis with a two-class subset of the MNIST dataset (LeCun et al., 1998). Since prior studies have shown that users prefer t-distributed stochastic neighbor embedding (t-SNE) over principal component analysis (PCA) and multidimensional scaling (MDS) for interactive visual labelling, t-SNE (van der Maaten and Hinton, 2008) is used for the similarity map.

For SPLOM, all bivariate combinations are shown in a matrix, and the user can select any of them to examine more closely in the scatterplot view. In the parallel coordinates view, the analyst can rearrange or invert dimensions and filter out records.

In general use, mVis allows the analyst to select one or multiple instances for labelling. Every time a set of instances are labelled, the Weka (a data mining software) implementation of a random forest classifier (Hall et al., 2009) runs in the background and suggests potential labels for all currently unlabelled instances by colour-coding according to their suggested class. Instances whose labels have been confirmed by the user are made visually distinct from instances with labels suggested by the classifier. Confirmed instances are shown as crosses in the scatterplots and similarity map and as thick lines in the parallel coordinates. Suggestions are shown as solid circles in the scatterplots and similarity map and as thin lines in the parallel coordinates. For the experiment described in this paper, the user was restricted to selecting a single instance at each step, which was then assigned its pre-assigned class.

Later, to assess the classification performance of the interactive visual labelling techniques, three methods were used: active learning, greedy selection,

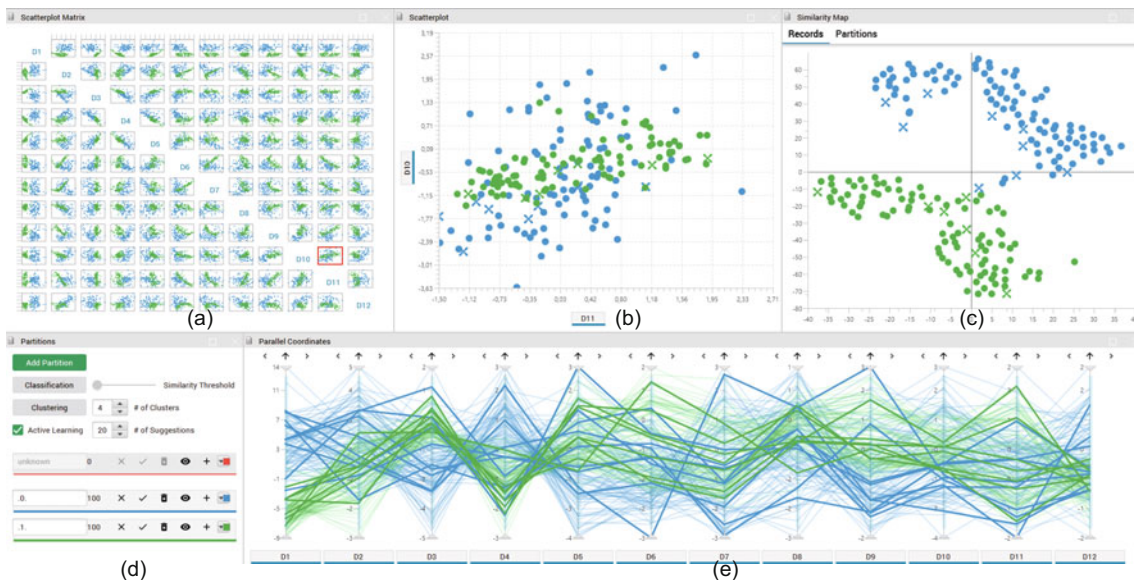


Fig. 1 mVis tool (Chegini et al., 2019a): (a) SPLOM; (b) detailed scatterplot; (c) similarity map; (d) parallel coordinates; (e) currently defined classes

Using the MNIST2 dataset, instances are colour-coded by class, here green for 1's and blue for 0's. Instances with confirmed labels are shown as crosses in the scatterplots and similarity map and as thick lines in the parallel coordinates. Suggestions from the classifier are shown as solid circles in the scatterplots and similarity map and as thin lines in the parallel coordinates. References to color refer to the online version of this figure

and random selection. Three active learning methods were used, smallest margin, entropy-based sampling, and least significant confidence, and the average accuracy in each step was used to compare the results. For the greedy method, the classifier was run for all possible instances for labelling, and the one with the best accuracy was selected. Greedy selection represents the best possible labelling result, and is the theoretical upper limit of what could be achieved by any visual labelling technique or active learning strategy. The random selection of instances was run 200 times, and the average accuracy in each step was used to compare the results. Random selection represents a practical lower limit for the accuracy that a classifier should achieve.

The work of Bernard et al. (2018a) was chosen to describe the strategies of users for selecting labelling candidates. Selection strategies were first grouped into data- and model-centred strategies. Data-centred strategies focus on the characteristics of data instances and include dense areas first, centroid first, equal spread, cluster borders, outliers, and ideal label. Model-centred strategies rely on visual feedback of the current state of the classification model and include class distribution minimisation, class borders, class intersection, and class outliers. In addition to the strategies defined by Bernard et al. (2018a), in this paper, another strategy was observed, which was named visual centre. Here, users would select instances in the centre of the visualisation on which they were currently focused.

4 Study design

A comparative experiment was conducted to evaluate the effectiveness of three individual visualisation techniques for interactive labelling, based on which records were selected by test users for labelling. The three techniques were similarity map, SPLOM with scatterplot for a detailed view, and parallel coordinates. The comparison was both quantitative and qualitative.

4.1 Research questions

The study addressed four main research questions:

RQ1: How do three individual visualisation techniques (similarity map, SPLOM, and parallel coordinates) compare in terms of accuracy of the

resulting classifier?

RQ2: How does interactive visual labelling with the three visualisation techniques compare to non-interactive labelling based on AL selection?

RQ3: Which of the three visualisation techniques are rated higher by users in terms of user experience and confidence during the selection of records to label?

RQ4: Do users adopt different labelling strategies depending on the visualisation being used?

The user was asked to choose 30 instances for labelling, one instance at a time, each of which was then labelled with its (correct) pre-assigned label from the ground truth.

Regarding RQ1, the accuracy of the classifier was computed each time after an instance had been chosen for labelling using the current training set (i.e., the set of records with confirmed labels at a particular point in time). The accuracy is simply the number of correct predictions divided by the total number of predictions. This experiment was concerned with which instances users chose to label, not with the actual labels which were then assigned. Hence, users were not actually asked to assign a label, simply to confirm the correct label from the ground truth (Fig. 2). To this end, after a user had chosen an instance to label, a pop-up window appeared showing the (pre-assigned) label for that instance, and the user was simply asked for confirmation. Once the label had been confirmed, the classifier ran in the background to refresh the suggested labels for currently labelled instances. Participants were provided neither with guidance nor with any AL suggestions about which instance to label next, but were asked to choose freely without time constraints. Participants were also not informed about the accuracy of the model as they worked, but they were shown a chart about accuracy after they had finished working with each dataset.

Regarding RQ2, the three active learning algorithms were run for each dataset, and the accuracy of the resulting classifier was calculated for each step, and then averaged over all three AL algorithms. This provides the baseline for comparison. The ratings for RQ3 were collected after the three visualisation had been investigated for each dataset. The labelling strategies used by each user for RQ4 were determined by analysing the thinking aloud protocol, screen recording, and interview responses.

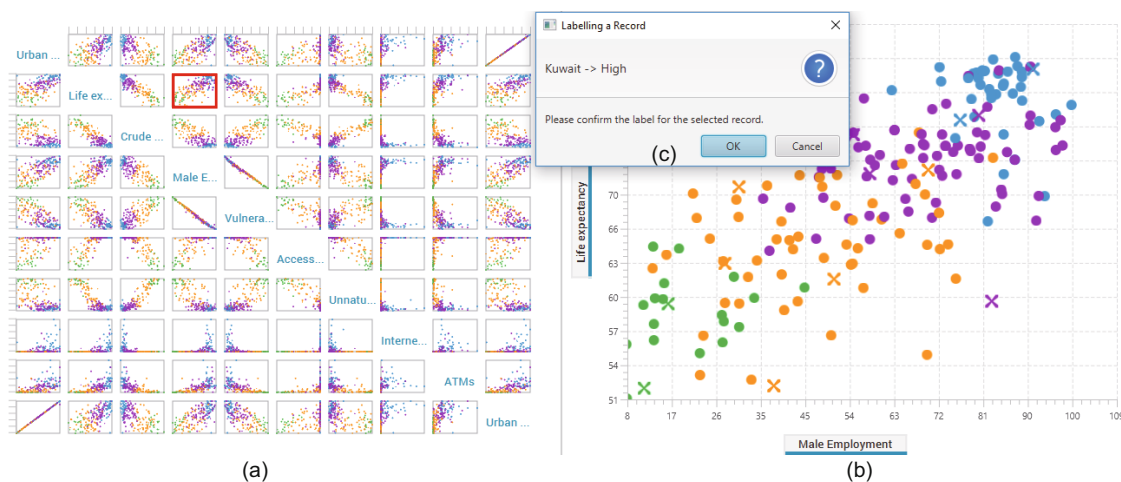


Fig. 2 SPLOM with scatterplot visualisation of the WB dataset used by a test participant

Instances are colour-coded by class. Instances with confirmed labels are shown as crosses, and suggestions from the classifier are shown as solid circles. The user has selected the scatterplot of life expectancy vs. male employment (red box) in the SPLOM (a) and has selected the instance of Kuwait for labelling in the detailed scatterplot view (b). The dialogue (c) asks the user to confirm the label for that instance. References to color refer to the online version of this figure

4.2 Datasets

Three datasets were used in this study. The first dataset is a two-class subset of the classic MNIST dataset (LeCun et al., 1998), comprising images of hand-written digits in one of two classes: 0 and 1. It was used to explain mVis to the participants in the tutorials phase of east test session. The 784 dimensions of the original dataset were reduced to 12 by PCA (Jolliffe, 2002) and named D1 through D12. The test dataset comprised 200 records with 100 records in each class. This dataset will be referred to as the MNIST2 dataset and is shown in Fig. 1.

The second dataset is an MNIST dataset with 50 records in each of four classes (200 records in total), representing digits 0, 1, 6, and 7. Like the first dataset, this dataset was reduced to 12 dimensions with PCA. This dataset will be referred to as the MNIST4 dataset. Figs. 3a and 4a show this dataset in parallel coordinates and a similarity map, respectively.

The third dataset is a socio-economic statistical dataset published by the World Bank (<https://data.worldbank.org/country>). Each record is a country. The 10 dimensions represent attributes such as urban population, life expectancy, access to electricity, and so on. The 192 records (countries) are unevenly classified into one of four economic classes: lower income, lower-middle income, upper-middle

income, and high income. This dataset will be referred to as the WB dataset. Figs. 2, 3b, and 4b show this dataset in SPLOM with scatterplot, parallel coordinates, and a similarity map, respectively.

4.3 Participants and setup

The study was carried out in a quiet lab. Ten participants were initially recruited for the study, but one was later eliminated from the analysis due to technical problems. Among the nine remaining

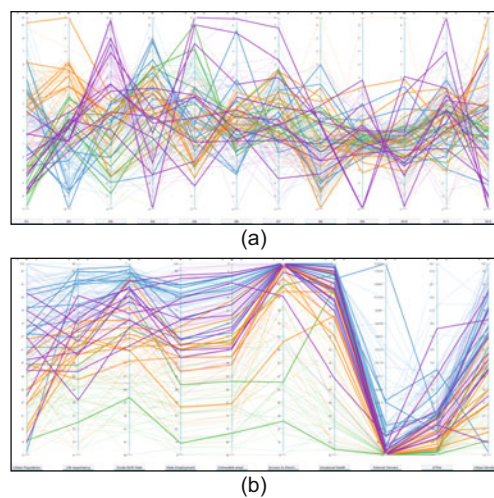


Fig. 3 Parallel coordinates visualisations of the MNIST4 (a) and WB (b) datasets

Instances are colour-coded by class. Instances with confirmed labels are shown as thick lines, and suggestions from the classifier are shown as thin lines

participants, three were female and six were male with a median age of 29 years old. All participants were familiar with machine learning and scatterplot visualisations. Two-thirds (6 of 9) were familiar with SPLOM and parallel coordinates. Two-thirds (a different 6 of 9) had previous experience in labelling multivariate datasets.

During their test session, participants were asked to think aloud, and to ask questions if they experienced any difficulties. At the end of this session, participants were encouraged to make suggestions for improvement. On average, each session lasted around 55 min, with the shortest and longest being 43 and 78 min, respectively. All sessions were captured by screen recording, and three sessions were additionally recorded with an external video camera for later analysis.

4.4 Procedure and tasks

The test procedure with each participant comprised four phases:

1. Opening: introduction and background questionnaire.
2. Tutorial: demonstration of mVis and practice with the MNIST2 dataset.
3. Test session: six experimental conditions, labelling each of the two datasets with each of the three visualisations.
4. Closing: interview with the participants.

In the first phase, the facilitator explained the purpose of the study and the participants then filled out a background questionnaire. The questionnaire included four binary (yes/no) questions. In these questions, it was asked whether the participant had used machine learning algorithms, scatterplots, SPLOM, and parallel coordinates.

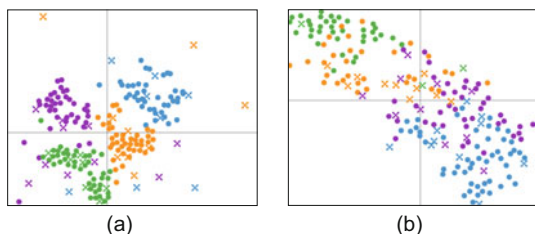


Fig. 4 Similarity maps of the MNIST4 (a) and WB (b) datasets

Different colours indicate different classes. Instances with confirmed labels are shown as crosses, and suggestions by the classifier are shown as solid circles. References to color refer to the online version of this figure

In the second phase, the facilitator first demonstrated the functionality of mVis with the MNIST2 dataset, explaining each of the three visualisation techniques and labelling two of the records. Then, users were asked to label a further 28 records using all three visualisations.

In the third phase, each test user performed the labelling task for each of the two datasets (MNIST4 and WB) with each of the three visualisations (similarity map, SPLOM with scatterplot, and parallel coordinates). Each visualisation was maximised to full screen. The presentation order of these six experimental conditions was grouped by the dataset but otherwise counterbalanced, as shown in Table 1. In each experimental condition, the test participant was asked to choose 30 instances for labelling (one after the other), which were then assigned their pre-assigned label (class). The experimental conditions were grouped by the dataset. One dataset was loaded, and labelling was completed with the three visualisations. Then the second dataset was loaded for the final three visualisations. After each dataset had been explored with all three visualisations, test participants were asked to rate their experience and confidence in labelling the records for each visualisation: (Q1) From 1 to 5, how do you rate the labelling experience with the visualisation technique? (Q2) From 1 to 5, how confident were you when selecting a new record with the visualisation technique? (1 was the worst and 5 the best rating) In Q1, it was clarified to participants to rate the experience of interactive labelling and not the ease of the user interface or other aspects.

Finally, in the fourth phase, the facilitator interviewed the test participants about their experience

Table 1 Presentation order of experimental conditions

Test participant	Experimental condition					
	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆
TP ₁	M-S	M-X	M-P	W-S	W-X	W-P
TP ₂	M-S	M-P	M-X	W-S	W-P	W-X
TP ₃	M-X	M-S	M-P	W-X	W-S	W-P
TP ₄	W-P	W-S	W-X	M-P	M-S	M-X
TP ₅	W-P	W-X	W-S	M-P	M-X	M-S
TP ₆	W-X	W-P	W-S	M-X	M-P	M-S
TP ₇	M-X	M-P	M-S	W-X	W-P	W-S
TP ₈	W-X	W-S	W-P	M-X	M-S	M-P
TP ₉	M-P	M-S	M-X	W-P	W-S	W-X

M: MNIST4 dataset; W: WB dataset. S: similarity map; X: SPLOM with scatterplot; P: parallel coordinates

and encouraged them to offer any feedback or suggestions they might have.

5 Results

The results of the study will be discussed for each of the three visualisation techniques (similarity map, SPLOM with scatterplot, and parallel coordinates) in terms of the four research questions from Section 4.1.

5.1 Similarity map

In terms of accuracy (RQ1), the similarity map outperformed SPLOM with scatterplot and parallel coordinates when using both the MNIST4 and WB datasets (Fig. 5).

Compared with active learning (RQ2), the similarity map consistently outperformed active learning for both datasets (Fig. 6).

Regarding the ratings of users (RQ3), the similarity map was rated higher than the other two visualisation techniques, in terms of both labelling experience and selection confidence (Fig. 7). Indeed, for labelling experience with the MNIST4 dataset, the mean rating of the similarity map was statistically significantly higher than that of the other two visualisations. All other differences in mean ratings were not statistically significant.

For both rating questions, the similarity map was rated slightly higher for the MNIST4 dataset than for the WB dataset. The reason could be that the clusters in the MNIST4 dataset were more distinct and visible than those in the WB dataset, as shown in Fig. 4b. This problem persists even when the projection algorithm for the similarity map is changed from t-SNE to PCA or MDS (Kruskal, 1964).

When using the similarity map, the strategies used by participants (RQ4) were similar to strategies observed during previous studies (Bernard et al., 2018a). In the similarity map, users tended to find distinct clusters from the beginning using the centroid first strategy. Therefore, the similarity map technique suffered less from the bootstrap problem (Fig. 5). After identifying distinct clusters, users tried to find outliers and make clear borders. The second main strategy used by participants was class intersection, i.e., selecting records that are in the wrong visual section. These records are closer to

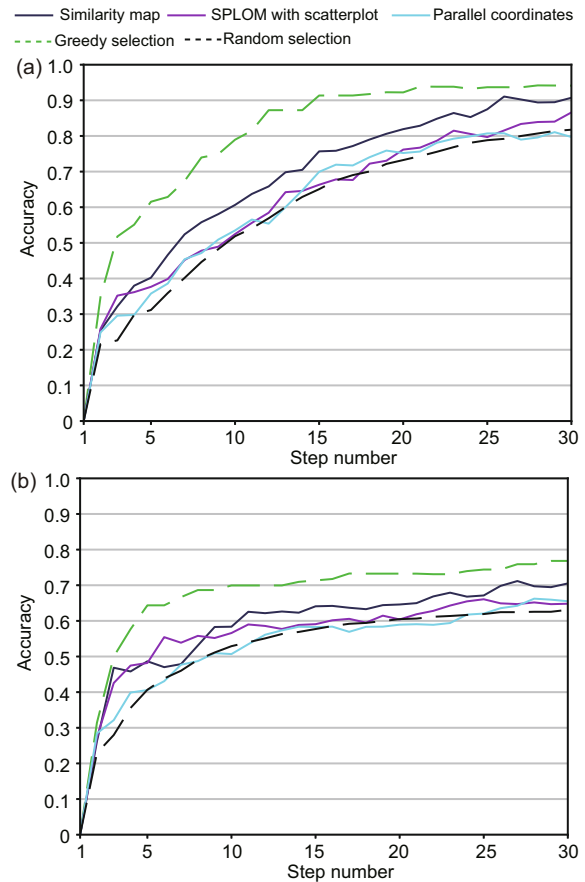


Fig. 5 Accuracy of visual labelling depending on the interactive visualisation techniques for the MNIST4 (a) and WB (b) datasets

The x -axis (step number) represents the cumulative number of instances already labelled. Greedy selection (green) represents a theoretical upper limit. Random selection (black) represents a practical lower limit. References to color refer to the online version of this figure

a different cluster than their own. Based on the observations, identifying suspected incorrectly labelled records in a similarity map was found to be a rather well-defined task by the participants. Note that the accuracy of these labelling strategies depends on the quality of the similarity map, e.g., how faithfully distances in the high-dimensional data space are preserved in the two-dimensional projection space. An interesting variant for a future experiment would be to include measures for projection quality in the similarity map, for which different visualisation techniques exist (Schreck et al., 2010).

5.2 SPLOM with scatterplot

Regarding the accuracy of the technique (RQ1), SPLOM with scatterplot performed slightly worse than the similarity map with both datasets, but

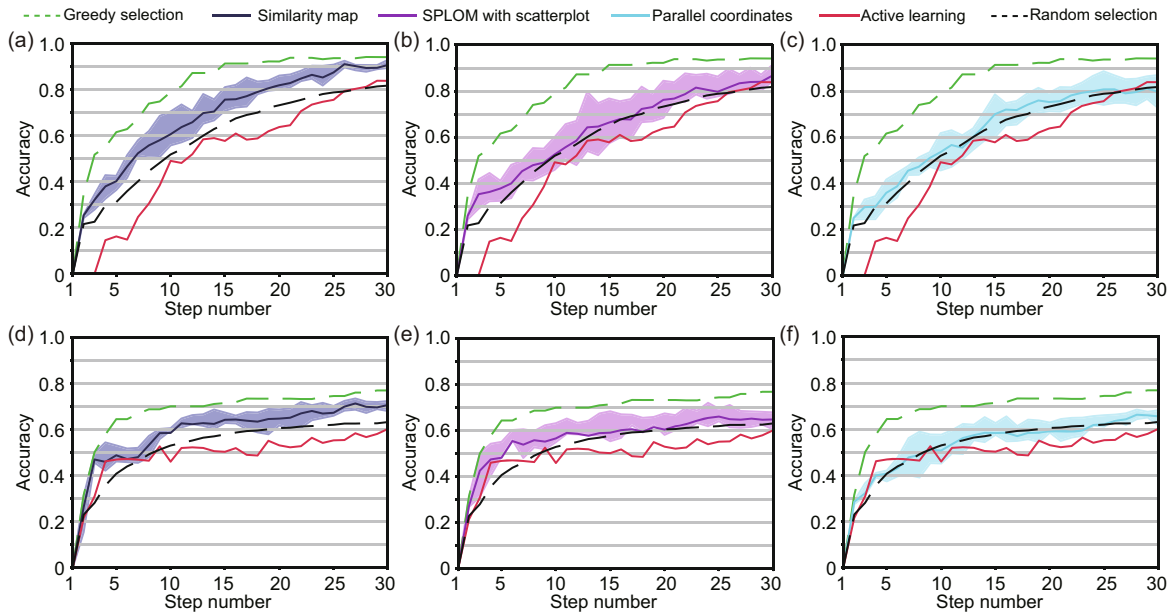


Fig. 6 Accuracy of the three interactive visual labelling techniques compared with active learning (red lines) for the MNIST4 and WB datasets: (a) similarity map (MNIST4); (b) SPLOM with scatterplot (MNIST4); (c) parallel coordinates (MNIST4); (d) similarity map (WB); (e) SPLOM with scatterplot (WB); (f) parallel coordinates (WB)

The semi-transparent coloured areas show the 25% and 75% quartiles. References to color refer to the online version of this figure

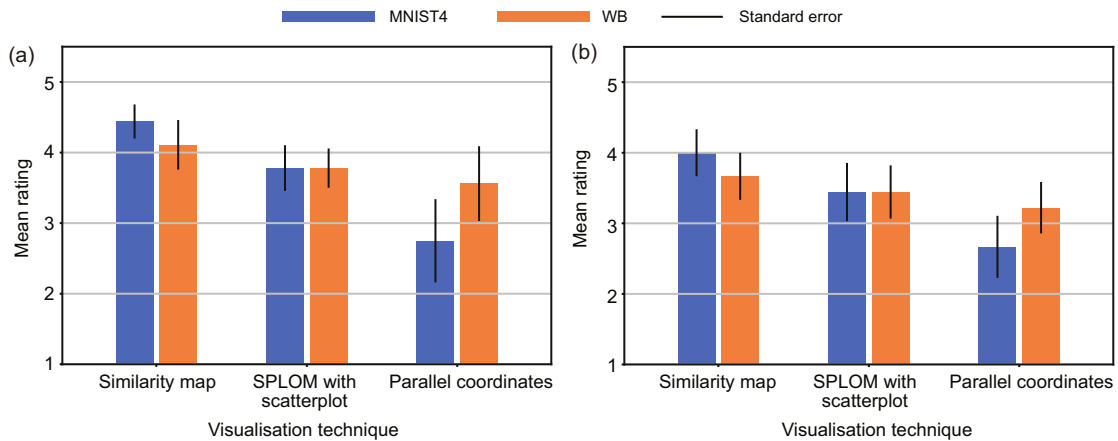


Fig. 7 Mean ratings of labelling experience (a) and selection confidence (b) for each of the three visualisations on a scale of 1 (worst) to 5 (best)

slightly better than parallel coordinates with the MNIST4 dataset and similar to parallel coordinates with the WB dataset (Fig. 5). The advantage of SPLOM compared to similarity map and parallel coordinates is that it suffered less from the bootstrap problem.

SPLOM with scatterplot outperformed the AL techniques (RQ2) for both datasets (Fig. 6).

Regarding the ratings of users (RQ3), the SPLOM with scatterplot technique was rated slightly lower than similarity map and slightly higher than parallel coordinates for both rating questions and with both datasets (Fig. 7). However, the only statistically significant difference is the lower mean rating for labelling experience of SPLOM with scatterplot compared to that of similarity map with the MNIST4

dataset. Regardless of the ratings for both datasets being similar, users stated that selecting candidates with the WB dataset was easier, since the dimension names were semantically meaningful and therefore more understandable.

Regarding the labelling strategy (RQ4), when using the SPLOM with scatterplot technique, users first attempted to find a scatterplot with well-spread records and then used the centroid first strategy on this scatterplot. Later, some users selected scatterplots with well-separated clusters. Others preferred to select scatterplots, which lacked well-separated clusters, and attempted to separate them. Some users tried brushing and linking to find outliers. Most users tended to select a single scatterplot and continued to use it instead of changing to a different scatterplot. With the MNIST4 dataset, which lacks semantically meaningful dimensions, users selected a scatterplot with a clearer visual pattern, for example, linear. Furthermore, users often selected scatterplots located in the centre of the SPLOM and ignored those in the outer reaches.

In general, users select scatterplots from SPLOM in the following order: (1) Scatterplots have a specific pattern (for example, linear); (2) Scatterplots have well-separated classes; (3) Scatterplots have overlapping classes; (4) If the dimensions have semantically meaningful labels, users select an interesting pair of dimensions based on the context; (5) Users randomly select scatterplots located in the centre of SPLOM.

The disadvantage of the SPLOM with scatterplot technique is that it has many false positives. That is, clusters were not always visible and well separated, which confused some users. Moreover, the SPLOM technique was sometimes overwhelming for users.

5.3 Parallel coordinates

Understanding parallel coordinates was hard for the users, mainly due to their lack of experience with this technique. Participants who were familiar with parallel coordinates performed better and were more confident during the experiment. Identifying patterns was difficult, particularly with the MNIST4 dataset. Furthermore, parallel coordinates tended to be more cluttered, and therefore selections became more random over time. Some users were frustrated when they were forced to select points from parallel

coordinates. On the one hand, one of the advantages of parallel coordinates was that it well guided the user's visual attention to extremes (peaks and valleys), enabling the users to easily identify these values. Furthermore, when users attempted to make borders for clusters in one single axis, using parallel coordinates was beneficial. On the other hand, one disadvantage of parallel coordinates was its lack of visual feedback, as stated by some users. Moreover, since users often focused on the centre of visualisation, the ordering of the axes was important when using parallel coordinates. Observations showed that if users rearranged the order of the axes, their experience could improve.

Regarding the accuracy of the classifier (RQ1), parallel coordinates performed as poorly as SPLOM with scatterplot with the MNIST4 dataset and slightly worse than SPLOM with scatterplot with the WB dataset. Parallel coordinates also suffered from the bootstrap problem, due to the users' tendency to select extreme values (peak and valleys) in the beginning and ignore the middle records, which usually included lower-middle income and upper-middle income countries.

Parallel coordinates outperformed AL (RQ2) in both datasets, although AL caught up as more instances were labelled (Fig. 6).

Regarding user ratings (RQ3), parallel coordinates received the lowest ratings, in terms of both labelling experience and selection confidence for both datasets (Fig. 7). The only statistically significant difference is the much lower mean rating for labelling experience for parallel coordinates compared to that for the similarity map with the MNIST4 dataset. However, the mean can be misleading. Half of the users rated parallel coordinates 5 when applied to the WB dataset, while the other half rated it poorly. The observations and interviews confirmed that some users strongly preferred parallel coordinates when the clusters were well separated, whilst others favoured other techniques.

In terms of the labelling strategy (RQ4), participants carried out the following strategies when using parallel coordinates: (1) selecting records on a single axis based on the values, (2) focusing on a combination of two axes, i.e., a line, (3) focusing on the shape of the polyline or general picture in three or more axes, (4) focusing on peaks and valleys, and (5) randomly selecting records on one axis or on a

line between two axes. The users' main strategy was to select extreme values in an axis located in the centre of the visualisation. The density first strategy was a common strategy used by the participants. At the beginning of the tasks, 60% of the participants used the default order of the axes, and 40% customised the order (mVis allows to reorder axes interactively). Users rarely changed the order of the axes afterwards. When using parallel coordinates, users paid less attention to having an equal spread strategy, and therefore the clusters were more imbalanced. Users also tried to identify class borders; however, finding such borders was difficult in the MNIST4 dataset.

When using the parallel coordinates technique, some users occasionally became frustrated and selected random records located in the visual centre of the plots. A recurring problem was the users' tendency to select outliers, leading to the bootstrap problem, as shown in Fig. 5. Furthermore, users selected higher values (peaks) more than lower values (valleys), which leads to an imbalance in the selection of peaks and valleys. When using parallel coordinates, users deployed the ideal labels strategy more than when using other techniques.

6 Discussion

The results of the study are promising as they show that the classification performance of interactive visual labelling techniques can outperform those of AL selection strategies. As shown in Fig. 6, with the WB dataset, all three visualisations perform better after around 10 labelled instances than AL. In contrast, with the MNIST4 dataset, AL catches up with the interactive visual labelling techniques as the number of labelled instances increases.

Only a very limited number of test users participated in this study. The experiments would need to be repeated with a much larger number of test users, in order for the results to be more generalisable. The results were also obtained for very specific choices of visualisation and datasets, and their generalisation would require additional validation. Labelling a dataset can be a dull task. Three participants mentioned that interactive visual labelling is enjoyable and feels like playing a game.

Some visualisations appear to be better suited to interactive visual labelling than others. The

similarity map seems to be the preferred view for labelling. This can be attributed to the fact that the similarity map reduces data, gives an overview of the similarity relationship, and is less complex than SPLOM with scatterplot or parallel coordinates. However, it is observed that when some example labelling is already available, some users prefer to use SPLOM with scatterplot for a more detailed insight into the high-dimensional data space and for label selection. It is also observed that users who are familiar with parallel coordinates perform better and are more confident in using it for label decision making.

Parallel coordinates and SPLOM are suitable for finding relationships between dimensions, identifying clusters, and exploring data to make sense of them. Visualisation of the labelled data in parallel coordinates could be improved. As the number of labelled instances increases, it can become overwhelming for the user to find the next instance to label. A problem found in all three visualisation techniques is false labelling. When an instance is close to a specific cluster, the user believes the instance belongs to that cluster and does not select it for labelling.

Regarding differences in the two datasets, it was observed that the MNIST4 dataset appeared very cluttered in the parallel coordinates visualisation, and patterns were difficult to discern. Therefore, the results for this test condition may have suffered. In the WB dataset, the dimensions had semantically meaningful names, and users felt more comfortable when choosing the axes in the parallel coordinates visualisation and choosing a particular scatterplot from SPLOM. For example, users often chose the access to electricity axis for labelling low-income countries.

It is interesting to observe in Fig. 6 that active learning often performs worse than random selection, at least in terms of the simple metric of model accuracy. However, this study only looked at the first 30 labelled instances, and AL strategies often start poorly (because of the bootstrap problem) but outperform random selection in later phases (Attenberg and Provost, 2010; Kottke et al., 2017).

In terms of improvements, one user mentioned a lack of control over the arrangement of scatterplots within SPLOM. Another user mentioned that parallel coordinates and SPLOM might be adapted to show the most "important" dimensions. Such an

idea is presented in other work using eye-tracking (Chegini et al., 2019c). AL was also mentioned by a participant as an additional form of visual guidance (Ceneda et al., 2016) for visualisation techniques. Another participant was curious to see the accuracy of the classifier after the selection of every instance, together with the number of already labelled instances from each class.

7 Limitations and future work

While the findings of this study are interesting, they also depend on a number of choices made and require further analysis. For the experiments, a number of settings were fixed, which could be varied as well. Three specific visualisations (similarity map, SPLOM with scatterplot, and parallel coordinates) were chosen and these were used individually for the labelling task. Many visual analytics systems provide multiple linked views and dynamic brushing. Indeed, mVis provides these features too, but they were not used in this study to simplify its design. Multiple linked views and brushing could possibly mitigate some of the disadvantages of single techniques, and lead to a hierarchical selection strategy. For example, users might want to select a group of points as labelling candidates from the similarity view, and then switch to SPLOM or parallel coordinates for detailed selection and labelling. In the future, support might be included for automatic ordering of dimensions in parallel coordinates or arrangement of the plots in SPLOM. To compare classification performance, three different AL algorithms (smallest margin, entropy-based sampling, and least significant confidence) were selected. While the selected algorithms are robust and applicable for different classifiers, the design space of active labelling is large and more comparisons could be made.

In the accuracy comparison experiments, it is assumed that the user always assigns the true (ground truth) label for a data point, once it has been identified for labelling. While this corresponds to the notion of a user being an “oracle” in AL, labelling errors could also be considered in a future experiment. Users could be allowed to freely pick a label, or even introduce a new label during interactive visual labelling. This would increase experimental complexity, but allow even more realistic assessments.

In many practical situations, the type and number of labels are not known in advance, but are determined in an iterative process. Also, in many practical problems, high-dimensional data attributes are often complemented with additional metadata and background information. For example, countries in the WB dataset could be presented as map views. Including visualisation of such additional data and studying how it is used during the labelling process would be an interesting experiment to do.

In the future, it would be interesting to study the dynamics of the labelling process. For example, when the choice of labels changes over time, are there learning effects during labelling? In the experiment described in this paper, the number of labels was fixed at 30. A future experiment could let the user decide when to stop the labelling process. To this end, feature and model space visualisations could be helpful for the user to assess when label saturation has been reached.

8 Concluding remarks

This paper presents a study comparing three interactive visualisations with each other and with active learning for the purpose of labelling a multivariate dataset. This study also explored subjective user ratings for the three interactive visualisations and discussed the labelling strategies employed by users with each of them.

All three interactive visualisations performed better than active learning algorithms, in terms of classification accuracy (assuming that the user always assigns the correct label to a selected data instance). The similarity map performed better than both SPLOM with scatterplot and parallel coordinates in both the MNIST4 and WB datasets. Nevertheless, parallel coordinates and SPLOM with scatterplot are useful in their own right, especially for datasets where the dimensions have semantically meaningful names. The results support the view that a user-in-the-loop approach is beneficial for creating training datasets. Finally, this paper presents some ideas for future work and further studies.

Contributors

Mohammad CHEGINI implemented the mVis system and code necessary for the conduction of the study. Mohammad CHEGINI and Jürgen BERNARD designed

the study. Mohammad CHEGINI and Fatemeh CHEGINI drafted the manuscript. Jian CUI helped conduct the experiment and data processing. Alexei SOURIN, Keith ANDREWS, and Tobias SCHRECK contributed to the definition of the underlying research questions, and they revised and finalized the manuscript.

Compliance with ethics guidelines

Mohammad CHEGINI, Jürgen BERNARD, Jian CUI, Fatemeh CHEGINI, Alexei SOURIN, Keith ANDREWS, and Tobias SCHRECK declare that they have no conflict of interest.

References

- Attenberg J, Provost F, 2010. Inactive learning?: difficulties employing active learning in practice. *ACM SIGKDD Explor Newslett*, 12(2):36-41. <https://doi.org/10.1145/1964897.1964906>
- Bernard J, Hutter M, Zeppelzauer M, et al., 2018a. Comparing visual-interactive labeling with active learning: an experimental study. *IEEE Trans Vis Comput Graph*, 24(1):298-308. <https://doi.org/10.1109/TVCG.2017.2744818>
- Bernard J, Zeppelzauer M, Lehmann M, et al., 2018b. Towards user-centered active learning algorithms. *Comput Graph Forum*, 37(3):121-132. <https://doi.org/10.1111/cgf.13406>
- Bernard J, Zeppelzauer M, Sedlmair M, et al., 2018c. VIAL: a unified process for visual interactive labeling. *Vis Comput*, 34(9):1189-1207. <https://doi.org/10.1007/s00371-018-1500-3>
- Bishop CM, 2006. *Pattern Recognition and Machine Learning*. Springer, Berlin, Germany.
- Ceneda D, Gschwandtner T, May T, et al., 2016. Characterizing guidance in visual analytics. *IEEE Trans Vis Comput Graph*, 23(1):111-120. <https://doi.org/10.1109/TVCG.2016.2598468>
- Chegini M, Shao L, Gregor R, et al., 2018. Interactive visual exploration of local patterns in large scatterplot spaces. *Comput Graph Forum*, 37(3):99-109. <https://doi.org/10.1111/cgf.13404>
- Chegini M, Bernard J, Berger P, et al., 2019a. Interactive labelling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active learning. *Vis Inform*, 3(1):9-17. <https://doi.org/10.1016/j.visinf.2019.03.002>
- Chegini M, Bernard J, Shao L, et al., 2019b. mVis in the wild: pre-study of an interactive visual machine learning system for labelling. *IEEE Vis 2019 Workshop on Evaluation of Interactive Visual Machine Learning Systems*, p.1-4.
- Chegini M, Sourin A, Andrews K, et al., 2019c. Eye-tracking based adaptive parallel coordinates. *12th ACM SIGGRAPH Conf and Exhibition on Computer Graphics and Interactive Techniques in Asia*, Article 44. <https://doi.org/10.1145/3355056.3364563>
- Culotta A, McCallum A, 2005. Reducing labeling effort for structured prediction tasks. *National Conf on Artificial Intelligence*, p.746-751.
- Hall M, Frank E, Holmes G, et al., 2009. The weka data mining software: an update. *ACM SIGKDD Explor Newslett*, 11(1):10-18. <https://doi.org/10.1145/1656274.1656278>
- Heimerl F, Koch S, Bosch H, et al., 2012. Visual classifier training for text document retrieval. *IEEE Trans Vis Comput Graph*, 18(12):2839-2848. <https://doi.org/10.1109/TVCG.2012.277>
- Ho TK, 1995. Random decision forests. *3rd Int Conf on Document Analysis and Recognition*, p.278-282. <https://doi.org/10.1109/ICDAR.1995.598994>
- Höferlin B, Netzel R, Höferlin M, et al., 2012. Inter-active learning of ad-hoc classifiers for video visual analytics. *IEEE Conf on Visual Analytics Science and Technology*, p.23-32. <https://doi.org/10.1109/VAST.2012.6400492>
- Inselberg A, 1985. The plane with parallel coordinates. *Vis Comput*, 1(2):69-91. <https://doi.org/10.1007/BF01898350>
- Jolliffe I, 2002. *Principal Component Analysis*. Springer, New York, USA.
- Kottke D, Calma A, Huseljic D, et al., 2017. Challenges of reliable, realistic and comparable active learning evaluation. *Proc Interactive Adaptive Learning Workshop*, p.1-14.
- Kruskal JB, 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1-27. <https://doi.org/10.1007/BF02289565>
- LeCun Y, Bottou L, Bengio Y, et al., 1998. Gradient-based learning applied to document recognition. *Proc IEEE*, 86(11):2278-2324. <https://doi.org/10.1109/5.726791>
- van der Maaten L, Hinton G, 2008. Visualizing data using t-SNE. *J Mach Learn Res*, 9(2018):2579-2605.
- Scheffer T, Decomain C, Wrobel S, 2001. Active hidden Markov models for information extraction. *Int Conf on Advances in Intelligent Data Analysis*, p.309-318.
- Schreck T, von Landesberger T, Bremm S, 2010. Techniques for precision-based visual analysis of projected data. *Inform Vis*, 9(3):181-193. <https://doi.org/10.1057/ivs.2010.2>
- Settles B, 2009. *Active learning literature survey*. Technical Report No. 1648, Department of Computer Sciences, University of Wisconsin-Madison, WI, USA.
- Settles B, Craven M, 2008. An analysis of active learning strategies for sequence labeling tasks. *Proc Conf on Empirical Methods in Natural Language Processing*, p.1070-1079.
- Shao L, Mahajan A, Schreck T, et al., 2017. Interactive regression lens for exploring scatter plots. *Comput Graph Forum*, 36(3):157-166. <https://doi.org/10.1111/cgf.13176>
- Wu Y, Kozintsev I, Bouguet JY, et al., 2006. Sampling strategies for active learning in personal photo retrieval. *IEEE Int Conf on Multimedia and Expo*, p.529-532. <https://doi.org/10.1109/ICME.2006.262442>