

DELFI: Mislabelled Human Context Detection Using Multi-Feature Similarity Linking

Hamid Mansoor* Walter Gerych† Luke Buquicchio‡ Kevin Chandrasekaran§
Emmanuel Agu¶ Elke Rundensteiner||

Worcester Polytechnic Institute

ABSTRACT

Context Aware (CA) systems that adapt to user behaviors have many real-world uses. CA systems require accurately labeled training data to learn models of users' context behavior. Unfortunately, it is difficult to gather sufficient realistic context data in controlled environments where reliable labels can be gathered. Therefore, recent works have used *in-the-wild study* designs, where data is gathered through passive sensing devices such as smartphones while users periodically supply corresponding context labels. However, labels gathered this way can be unreliable as users may provide incomplete or inaccurate labels which makes it difficult to build robust CA models. We propose DELFI (Detecting Erroneous Labels using Feature-linking Insights), a visual analytics approach to discover and clean unlabeled or mislabeled context data. Visualizations enable highlighting of similar data to find patterns and anomalies in behaviors. However, this is challenging when working with erroneous human-labelled data as linking similar context labels is flawed since the labels themselves are in question. DELFI identifies probably-mislabeled instances by color-coding them based on an anomaly score. Additionally, DELFI links similar instances based on a novel concept called *Multi-Feature Similarity Linking*, which facilitates the identification of probably true labels of mislabeled and unlabeled data. We demonstrate the utility of our approach with use cases and evaluation from domain experts.

Index Terms: Human Context Data—Mislabelled Data—Interactive Data Visualizations—Unlabeled Data;

1 INTRODUCTION

Benefits of In-The-Wild Data Collection

Context aware systems that can accurately identify and adapt to their users' context have huge implications in multiple fields such as detecting a passenger's state in self-driving cars or accurately identifying medical symptoms early and triggering preventive measures in healthcare [3, 8]. However, it is impossible to collect realistic data on all possible contexts using a controlled study design [27]. For this reason, recent work in Context Recognition (CR) [22, 25] has focused on gathering real world data using *in-the-wild* data collection. This type of data is typically collected through sensor-equipped devices such as smartphones, where the device passively records some set of sensor values while the user periodically provides ground truth for their contexts which is then used to train CA algorithms [22, 25].

In-the-wild data collection is more feasible now than ever due to the proliferation of smartphones. Smartphones are useful for developing and deploying CA systems because they are widely used and are equipped with powerful sensors such as GPS, accelerometers, gyroscopes and microphones, which can collect data unobtrusively in the background while people perform their daily routines. Readings from these sensors can be used by analysts to identify the various contexts a person visits such as their location, body posture and activity. This objective data can then be classified using human-provided labels about the corresponding contexts as ground truth.

State of the Art Visual Analytics In The Wild Human Behavior Data

There has been interest in quantifying and understanding human behavior in uncontrolled environments (*in-the-wild*) such as micro-blogging sites and social media [13]. State of the art interactive visual analytics for *in-the-wild* user behavior typically highlight data with similar behavior labels [2, 14, 18, 28]. These visual approaches are powerful because human behavior is multifaceted and difficult to understand from basic summary statistics. Data visualizations enable analysts to highlight instances of similar data to find patterns and make inferences about behaviors. These visual solutions generally deal with data where context is verifiable, such as in the case of social media behavior or email logs where the user's behavior is known. For example, the number of re-tweets or the user's email response rate and content is verifiable.

Limitations of State of the Art for User-Labeled Data

State of the art human behavior visualizations are not robust enough to transfer to domains where the "ground truth" is provided by the users. This is due to the fact that highlighting instances with the same human contexts labels is flawed as the labels themselves are in question. This type of user-labeled data can arise in studies where users are expected to report their physical activities [23]. These studies are scalable and can be deployed to a larger and more diverse audience than any lab setting, but the "ground truth" context labels collected through this method are prone to error as these context labels are provided by the study participants themselves. Participants often do not have a strong incentive to provide clean labels, and are also prone to human error. This limits the robustness of any CR models trained on data gathered in-the-wild, as the ground truth these models are trained on is flawed.

As the ground truth of contexts in user-labeled data is unreliable, it is desirable to have a visual analytics approach that can highlight data based on an alternate similarity metric. As these flawed user-supplied labels are often used as ground truth to train machine learning classifiers [22], it is also desirable for a visual solution to identify instances that have probably been mislabeled by the user. Once identified, the analyst would also benefit from discovering the true context of that mislabeled instance. Furthermore, there are often many unlabeled instances in user-labeled human context data [22], where it would be beneficial for the analyst to know the

*e-mail: hmansoor@wpi.edu

†e-mail: wgerych@wpi.edu

‡e-mail: ljbquicchio@wpi.edu

§e-mail: kchandrasekaran@wpi.edu

¶e-mail: emmanuel@wpi.edu

||e-mail: rundenst@wpi.edu

true labels of these unlabeled instances.

Proposed Visual Solution for User-Labeled In-The-Wild Limitations

We present DELFI, an interactive data visualization and exploration tool to discover instances of mislabeled data and missing labels. DELFI is a visual analytics solution for user-labeled in-the-wild human context smartphone-collected data. DELFI identifies data instances that have probably been mislabeled by the user by calculating an anomaly score using the Isolation Forest [15] algorithm. This score gives a measure of how much a context differs in the feature space from other instances with the same context labels. The intuition is that mislabeled instances will be unlike the correctly-labeled instances of it's incorrectly assigned context labels.

Any anomaly detection method has limitations, especially when dealing with the typically high dimensional and multi-faceted nature of human context data. Rather than only relying on an anomaly score to decide mislabeled instances, DELFI also links highlights instances based on a *multi-feature similarity score*. This score is a novel metric we introduce, which measures how similar instances are based on their features, not labels. This highlighting enables the analysts to quickly identify what other instances a suspected-mislabeled instance is most similar to. If the most similar instances share the same context label as the suspected-mislabeled instance, the analyst can conclude that the instance was not in fact mislabeled. However, if the most similar context labels are consistently different from the provided label for that instance, the analyst is justified in believing that the instance was in fact mislabeled. Additionally, the analyst will be able to identify what the true label likely is, based on the labels of the most similar instances. Additionally, the multi-feature similarity score allows the analyst to discover the likely labels of unlabeled instances. That is, the most common labels of the most similar instances to a given unlabeled instance are likely to be the true labels of that instance.

Furthermore, since it is infeasible for the analysts to correct these labels on an instance-by-instance basis (as in such studies an instance might be a second or minute worth of data, where the total duration of the study might be weeks of data [22]), we group sequences of contexts with the same context labels into *continuous context chunks*, and provide the aforementioned visual analytics on these longer chunks.

In this paper, we thus make the following contributions:

- Research and development of DELFI, a visual analytics system to interactively explore human context data
- Demonstration of DELFI by utilizing use cases to show its utility in finding mislabeled instances and providing labels for unlabeled instances
- Propose a novel data instance similarity metric called *multi-feature similarity score*
- Propose a novel concept called “Multi-Feature Similarity Linking” that highlights data that is similar based on their feature values, not labels.
- Evaluate DELFI in sessions with experts in the field of human context modeling and present results.

2 RELATED WORKS

2.1 Visualizing Behavior Data To Find Trends and Patterns

People’s usage behaviors with various technologies in uncontrolled environments such as social media [7, 14] leave complex and interesting data trails that can reveal a lot of interesting relationships.

Archambault et.al. [2] proposed ThemeCrowds, a visualization system to interactively mine and discover trends in Twitter data. Their tool sets various levels of resolution for data mining to show the most important Twitter users for particular topics and the evolution of topics online. Setting various granularity levels is important in human context data because it typically contains a large number of events that will be difficult to understand without some grouping. Wang et.al. [26] used interactive visual techniques to present temporal summaries at various levels of temporal granularity to enable analysts to discover trends in timeline data. Showing such evolving trends is important in visualizing human context data so that analysts can quickly discern unlikely occurrences of activities.

Data visualization is also useful for individual vs. group analysis and making comparisons across various trends. Polack et.al. [20] present Chronodes, an interactive visualization tool to visualize mHealth (Mobile Health) sensor data and highlight patterns in various events. Chronodes also enabled users to define and compare groups of peoples’ behavior data. Such comparison of objective sensor data is necessary for human-labelled data as the labels themselves might not reflect reality.

Nguyen et.al. [18] created U4, an interactive tool to find instances of unusual behavior in event data using an online administrative tool. They proposed a novel concept called Multi-Semantic Linking, which proposes linking data based on semantic similarity across connected panes for more intuitive exploration. Highlighting such patterns and sequences across multiple views is especially important in human labelled context data as that allows analysts to gain clearer insight about the data labels that rarely occur in a user’s data and which may be indicative of faulty labelling.

2.2 Visualizing Behavior Data In Conjunction With Automated Anomaly Analysis

Visual analytics is an effective way to find anomalous points in crowdsourced data. Liu [16] et.al. created *LabelInspect*, an interactive visualization tool to help experts identify instances of wrongly annotated image data that was obtained through crowdsourcing. Their integrated visualization system allowed analysts to reduce “noise” in their labels and improve the image dataset. LabelInspect also identified suspicious workers and let the analyst make judgments about the quality of work they were receiving. Their work shows the utility of visual analytics in improving data quality when the annotations are in doubt.

Calculating measures of anomaly for human behavior data is difficult because of vast behavior variation and multivariate nature of the data. Therefore data visualizations can reduce the opacity of any black-box automated method to generate anomaly scores. Cao et.al. [5] created TargetVue, an interactive system that allows users to find anomalous behavior data in online communication systems. Their system presents rich user data using multiple glyphs and connected panes that make the task of human judgments easier. A common pitfall in automated analysis of anomalous online information data is that it fails to consider other *semantic* information that an analyst might use to explain the degree of “unusualness”. To tackle this, Zhao et.al. [28] created a visualization tool called FluxFlow to analyze twitter data. They used advanced machine learning models to find instances of anomalous tweets and then present them in linked panes for further analysis. Such approaches are particularly relevant because unlike some domains such as cyber security where user log data can be used as a ground truth, human labelled context data makes it challenging to establish a ground truth.

3 HUMAN BEHAVIORAL CONTEXT

A person’s context is difficult to detect [22]. Human context comprises of several factors such as a person’s location, activity, body posture, occupation, time of day etc. In order to collect realistic context data for the vast range of possible contexts a user

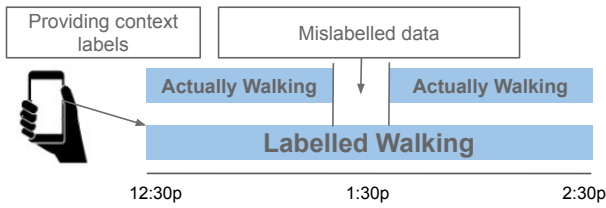


Figure 1: Typical In-The-Wild Labelling Workflow.

could be in, it is desirable to collect this data using an in-the-wild method. To enable this, Vaizman et.al. [23] created the *ExtraSensory* phone application which continuously gathered data from multiple smartphone sensors throughout the day and had an intuitive interface to enable users to label their contexts for different time periods periodically. Users typically provided multiple labels throughout their day.

Errors in Human Context Data

In-the-wild-gathered human context data is prone to having instances of mislabeled or unlabeled data, as users are typically required to label their own data. [6, 22, 25]. Generally, factors such as recall bias, careless reporting and inopportune solicitation of context labels [6] are reasons for why individuals often fail to properly label their own context data (Figure 1). From literature review, there seem to be two broad issues with the labels provided in in-the-wild studies:

- **E1** - Wrong labels: Users are prone to memory biases and may mis-estimate both the start/end times as well as the durations of their contexts. The users may also get interrupted at inopportune moments and may provide labels carelessly.
- **E2** - Missing labels: Users may not provide labels for all the collected sensor data or may provide incomplete data. For instance a user may be “Sitting”, “Typing”, with “Phone in their Hand” but they only supplied the context “Sitting”. Missing labels make it challenging to use the data for classification.

4 GATHERING CONTEXT DATA

4.1 WASHSensory

In a pilot study, we used a modified version of the *ExtraSensory* Android application and renamed it *WASHSensory*. This application has new context labels and collects additional phone data such as the apps being used in the foreground. The data was gathered to train classifiers on real world context data to determine patterns in behavior that may be indicative of traumatic brain injuries and infectious diseases, for instance determining if a person is using the bathroom more often or is spending less time lying down and sleeping. The data collection duration and the scheduling of the collection of sessions were not changed. The user labelling mechanisms were also kept the same. The application was installed on participants’ phones where it collected approximately twenty seconds of data between minute long intervals throughout the day. The participants were able to label the data either by “Actively” labelling what they will be doing in the near future or use a “History Page” which lets them report their activities retroactively. The users could also respond to notifications asking for labels. Table 1 shows the eighteen labels that they were able to provide. The users can only select one label for phone priception i.e. “Phone In Hand” vs “Phone In Pocket” and can select multiple other labels. The participants were asked not to modify their natural routines for the duration of the study.

Overall, 115000 labelled individual data sessions were gathered across twelve study participants. The average user participation duration was two weeks. An individual session consists of approximately twenty seconds of continuous and discreet sensor data along

Table 1: Context Labels that study participants could select on WASHSensory.

Context Labels		
Phone In bag	Phone In hand	Phone In pocket
Phone On Table - up	Phone On Table - down	Walking
Bathroom	Standing	Jogging
Exercising	Running	Sitting
Sleeping	Lying down	Typing
Going Up Stairs	Talking On Phone	Going Down Stairs

with phone state measurements between minute long intervals. The participants provided labels for approximately 65% of the data. The labels and the data are merged and processed so that every data session has summary features such as the averages and standard deviations of multiple sensors etc. In the end, we computed over 120 features for every data session. These feature values are used for training and testing of context recognition models.

4.2 Label Providing Mechanisms

An important aspect of gathering data in-the-wild is that it needs to be *unobtrusive* i.e. it should not require a substantial effort or disrupt a person’s daily routine, otherwise it cannot truly be considered “real world” data. That is why *ExtraSensory* and *WASHSensory* applications let the users provide labels for the near future and past (for the present day and yesterday using a “History Page”) and also let them set up notification schedules to ask for labels. There has been work that examines the effect of using various labelling mechanisms and the accuracy of the provided data. Chang et.al. [6] described the labelling process in terms of “User Balance” and “User Load”. They collected real world travel data and had the subjects provide labels for their travelling activities. They categorized three categories of annotation methods called “POST”, “PART” and “SITU” which correspond to labelling data post-hoc, labelling data for near future and labelling data in-situ respectively. Their work suggests that while the PART condition provides the lowest quantity of data, it provides the best quality data. *ExtraSensory* and *WASHSensory* also enable users to label data in ways that can broadly fit these categories. Since the labelling sources can have implications for the quality of data, a visual analytics approach may encode labelling mechanisms for the collected context labels which may aid analysts in making judgments about the authenticity of the labels.

5 CALCULATING OBJECTIVE MEASURES TO COMPARE DATA

5.1 Data With Similar Feature Values

People are prone to making errors in the labeling process. This means that the reliability of CA systems is hindered by unreliable ground truth labels. One way of identifying mislabeled instances is by removing data that is anomalous given its context labels. However, this method is limited by the effectiveness of the anomaly detection method. Additionally, a simple anomaly score is unable to identify the likely labels of unlabeled instances. For this reason, we compute an anomaly score to identify possible mislabeled instances and an additional metric which we call the *feature similarity score* to supplement the anomaly score. The goal of this similarity score is to determine which context a suspected-mislabeled session identified through an anomaly detection method is most similar to. If a suspected-mislabeled session is very similar to a set of context labels different from the labels that were supplied by the user, this informs the analyst that the session is likely truly mislabeled and gives the analyst insight into what the correct labels likely are. As there are typically hundreds of thousands of instances in human

context recognition datasets (as each instance typically represents data collected every second or minute, where the study might take place over weeks [22]), we compute these aforementioned metrics for linking the data on aggregated sequences of context data, where every instance in the sequence has the same context labels.

Thus, first We segment a users data into *Continuous Context Chunks*, then compute an *Anomaly Score* for each continuous sequence of context data in order to identify likely-misabeled instances, and then compute a *Multi-Feature Similarity Score* between every context chunk in order to supplement the anomaly score and give insight into the true context labels for mislabeled and unlabeled context chunks. Each step - segmenting the data, computing the anomaly score, and calculating the similarity score - is described below.

5.2 Continuous Context Chunk

Let D be a dataset of context data from N users, where each user's data consists of a sequence of feature and label pairs $U_i = \{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$, where $\|x_i\| = d$ is the number of features of each reading. $y_i \in \mathbb{R}^c \cup \{\emptyset\}$, as each label can either be a vector of length c where c is the cardinality of the set of all labels or can be empty if the user did not provide labels for that instance. Additionally, each (x_i, y_i) pair of feature and label values has an associated timestamp t_i .

We define a *continuous context chunk* to be a subsequence $C_j \subset U_i$, $C_j = \{(x_k, y_k), (x_l, y_l), \dots, (x_m, y_m)\}$ where $y_g = y_h \forall$ label pairs in C_j , $y_{k-1} \neq y_k$ and $y_m \neq y_{m+1}$, and the difference between associated subsequent timestamps t_i and t_{i+1} are less than some Δt for all subsequent feature-label pairs. That is, a continuous context chunk is a sequence of context labels and associated feature values where all the context labels in the sequence are equal, and the time difference between each instance is less than some cutoff Δt .

We chose $\Delta t = 300$ seconds for our work. This is due to the fact that the data collection mechanism may stop collecting data for some period of time. In that case, even if the context of the user before the data collection stopped matches the context of the user after data collection begins again we would not want to consider that to be one continuous context chunk. We overcome this by setting the required maximum time difference between contexts in a continuous context chunk to be relatively small, but still long enough to account for the longest systematic delay between data collections we found among the devices used during our study.

5.3 Anomaly Score

We now describe how the anomaly score for a given continuous context chunk is determined. As one of the goals of our method is to identify instances of mislabeling, if there was some measure of how dissimilar a given continuous activity chunk is compared to other continuous chunks with the same context-labels we could identify these mislabeled instances. This is due to our assumption that mislabeled instances will have feature values that are unlike the feature values of true instances of that context.

We use the anomaly score for a given instance as its measure of dissimilarity to other instances with the same context labels. We calculate the anomaly score from the features associated with each context instance, which can include both numeric and semantic features. The feature-values used in human activity recognition are typically high-dimensional [24]. For this reason it is critical for our anomaly detection method to be robust to high-dimensional data. For this reason, along with the fact that we can not assume that our feature values follow a normal distribution, common anomaly detection measures such as the z-score are not appropriate. Therefore, we selected the IsolationForest algorithm as our anomaly detection method. This is due to the fact that IsolationForest is robust to high-dimensional data,

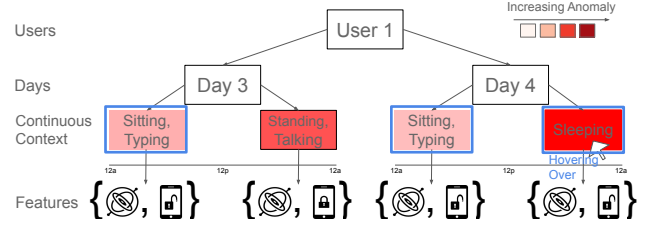


Figure 2: Multi-Feature Similarity Linking. Highlighting instances of context data that are *similar* in terms of *feature* values rather than in terms of annotation.

and does not assume a specific distribution of the feature values [15].

Isolation Forest

Isolation Forest is an anomaly detection method which isolates outliers, namely points with lower local densities than the majority of points. The method works by creating a set of decision trees, where each tree successively picks a dimension and value on which to split the data. Each tree repeats this step until every point is isolated. This assumes that outliers will be isolated in fewer splits than inliers on average. Mathematically, the decision function is:

$$s(x, n) = 2^{-\frac{E(g(x))}{c(n)}}$$

where $E(g(x))$ is the expected value of the path length $g(x)$ to isolate observation x , n the number of data points in the training set, and $c(n)$ the average path length to isolate an observation. A score close to 1 is classified as an outlier.

If $c(n) \ll E(g(x))$, namely if the average path length for an observation is significantly smaller than the expected path length for all data points, then that point will have a high anomaly score.

Anomaly Score for Continuous Context Chunks

Given an instance of data consisting of a feature-label pair (x_i, y_i) , $y_i \in \mathbb{R}^J$ in a continuous context chunk C , we compute an anomaly score $a_{i,j}$ for each positive label in the label vector y_i . We then average the anomaly score for each label over all instances of the context chunk.

5.4 Multi-Feature Similarity Linking

The anomaly score gives insight into which context chunks were likely mislabeled. However, it does not tell us what the true labels for that context chunk likely are. In order to supplement the anomaly score and identify the true labels of mislabeled and unlabeled chunks, it is desirable to find instances of data where the *feature values* are *similar* rather than viewing data based on identical labels. Knowing which other instances a given instance of context data is most similar to can aid the analyst in identifying and removing or relabeling mislabeled instances, as well as in labeling unlabeled instances.

Data visualizations allow analysts to gain insights by *linking* other instances of data with some related feature or value. Nguyen et.al. [18] proposed a novel visual concept called *Multi-Semantic Linking* that relaxed the constraints of normal linking behavior to highlight data across multiple panes and levels with similar *semantic* information. In the case of human labelled context data, the ground truth labels associated with the data are themselves in question. Therefore, Multi-Semantic Linking might not work in this case as any attempt to highlight the same *semantically related* contexts would need to assume completely accurate labelling which is not the case.

Inspired by Multi-Semantic Linking, we propose a novel visual concept called *Multi-Feature Similarity Linking*. The idea is to highlight and link data that is similar in terms of objective *features* since

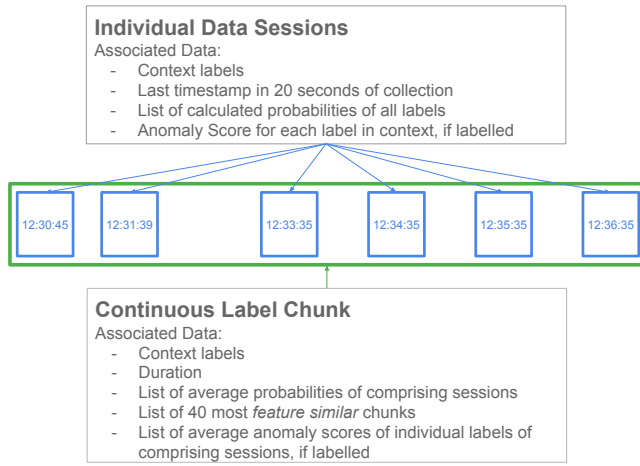




Figure 3: Data Description: Showing the various data attributes linked to individual data sessions and continuous context chunks.

the sensor data and other phone measurements may be indicative of a different context than the one labelled by the user. Figure 2 illustrates visually the concept of Multi-Feature Similarity Linking. In this example, the app user has provided labels for “Sitting” which through some anomaly metric stands out as being suspicious. If a user interacts with this data (in this case hovers over it), other instances of data where the feature values were similar are also highlighted. In this case the data for the “Standing” instance is close in terms of *features* but not labels to two instances of “Sitting, Typing”. For instance, the data instances with blue boxes around them have similar gyroscope  values and were also performed with the phone being unlocked . Further, the highlighted data for “Sitting, Typing” is also less anomalous. This may guide the analyst to make a more informed decision about mislabelled data. The metric we use for Multi-Feature Similarity Linking is a novel *multi-feature similarity score*.

5.5 Multi-Feature Similarity Score

For each instance continuous activity chunk K we associate a sequence of predicted context labels $P_K = \{\hat{y}_i, \hat{y}_j, \dots, \hat{y}_m\}$, $\hat{y}_i = f(x_i) = y_i + \epsilon_i$, where f is some classifier and ϵ_i is the difference between the prediction for the i th label and the value provided for that label.

For each continuous context periods K , we average the label probability vectors P_K . Let \bar{P}_K be this average, where $\bar{P}_K = \frac{\sum P_K}{|P_K|}$.

We then take Euclidean distance between continuous context chunk pairs K and J , for all such pairs. That is, we compute $S_{J,K} = d(\bar{P}_K, \bar{P}_J)$. $S_{J,K} = S_{K,J}$ is then the *feature similarity score* between continuous context chunks K and J .

Continuous context chunks whose associated label-probability vector have the smallest Euclidean distance with a given chunk are considered to be similar continuous context chunks.

5.6 Relabeling Unlabeled Sessions using the Feature Similarity Score

Data labeled in-the-wild generally contains a large number of unlabeled instances, where the user did not provide any context labels [22]. The predictions of a *non-standard classifier* that is trained to predict whether or not a label was provided for an instance can be transformed into the probability of an instance being a positive example of a given label under the assumption that the distribution of true-positives matches the distribution of unlabeled positives in the feature space [4]. Let Y_{true} be the set of true positives, and $Y_{unlabeled}$ be the set of unlabeled positive instances. Then,

$P(y_{unlabeled} = 1|x) = c * P(y_{true} = 1|x)$, where c is the labeled frequency $c = Pr(y_{unlabeled} = 1|y_{true} = 1)$. c is unknown in practice, but [9] shows that c can be estimated from the relative values of $P(Y_{unlabeled}|X)$. Thus, we can train a classifier f on both the labeled and unlabeled data, and then create the feature-similarity scores for all continuous chunks, including those that are unlabeled. The analyst can then examine which continuous chunks are most similar to the unlabeled chunk to estimate the true context label for that chunk, even if the classifier f would predict a value less than the positive-cutoff for that chunk.

5.7 Classification

In order to calculate the feature-similarity scores, we need to train a classifier f to predict the labels Y from data X . We chose f to be a *gradient boosting classifier* [10], as it has been shown to be a robust classifier that routinely shows high performance on a variety of classification problems [19]. Unlike a deep learning classifier, gradient boosting produces high classification accuracy on smaller datasets. As we perform in-user classification, and thus might have a limited number of samples, this is a desirable characteristic for our classifier.

5.8 Gradient Boosting Classifier

Given a dataset of features X and corresponding labels Y , we construct a *Gradient Boosted Classifier* f such that

$$f(X) = \sum_{i=0}^N h_i(X),$$

where h_i is a *weak learner*, which is typically chosen to be a shallow decision tree. Each individual learner h_i is trained to fit the residuals of the previous learners. That is,

$$h_i(X) = Y - \sum_{j=0}^{i-1} h_j(X)$$

The initial learner h_0 is trained to match the target Y .

6 DESIGNING A VISUAL APPROACH TO PRESENT HUMAN CONTEXT DATA

Interactive visualizations enables analysts explore and mine data to gather intuition about multi-variate and heterogenous data. Keim et.al. [12] discussed the challenges in presenting complex data using visual analytics and proposed the “*Visual Analytics Mantra*”:

- Analyse First (M1)
- Show the Important (M2)
- Zoom, Filter and Analyse Further (M3)
- Details on Demand (M4)

As mentioned, in-the-wild human-labelled context data is complex as it contains readings from a multitude of continuous sensors, along with other discreet measurements. This mantra was used to guide our workshops to design visualizations that would highlight mislabelled data. Given the limited visual real estate because of screen space and complexity of the data, there were a few specific visual requirements that had to be met in order to visualize this data effectively. They are summarized below:

- V1 - Grouping continuous instances of the same context to easily make sense of the data since there are 1440 minutes in a day and it would be difficult to analyze so many individual sessions separately (M1).



Figure 4: Detecting Erroneous Labels using Feature-linking Insights (DELFI). (A) *Habit View*: Shows every user’s data split up by days of participation. The bars represent Continuous Context Chunks. Hovering over a chunk hides all other chunks except for those that are most feature similar. The user has hovered over a chunk with labels “In pocket” (shortened from phone in pocket) and “Sitting”. (B) *Chunk Detail View*: Shows details about the clicked chunk such as the label providing mechanisms used to label the data sessions in the chunk. The labels in this chunk detail view were provided using the “History” interface. The individual labels comprising the context are split up as separate bars and the lines inside them show the respective anomaly score values or probability values (in case of unlabelled data) for the data sessions comprising the continuous context chunk. The two grey bars above the labels represent the battery charging status (top) and app usage status (second bar) at the time of collection for the data sessions. (C) *Relabel Dialog*: The analyst can relabel or unlabel data sessions by dragging the mouse across the label bars in a Chunk Detail View that they think are mislabelled and selecting a new label.

- V2 - Humans are creatures of habit so there needs to be a quick way to figure out patterns and see if anything is out of place (M2).
- V3 - An objective measure of anomaly along with what typical labels based on feature similarity would be helpful in guiding exploration (M3).
- V4 - Highlight data that have similar features while being *label agnostic* since the labels or the “ground truth” itself is in question (M3).
- V5 - Showing other phone state measurements apart from continuous sensor streams that can give clues about a person’s “ground truth” data. For instance, making sessions with a particular app conspicuous so the user can identify suspicious labels. i.e., if a person labels “Phone in Bag” and “Sleeping” but has a communications app open the foreground, that is suspicious (M4).

To allow for multi-perspective exploration of data, we present Detecting Erroneous Labels using Feature-linking Insights (DELFI), an interactive visual analytics tool to find instances of mislabelled human context data. We designed DELFI keeping in mind the issues with the dataset, the tasks that analysts need to perform to reveal those issues and the visual cues we want to highlight. To illustrate the features and use cases for this tool, we introduce Jill, a graduate student who wants to train CR models and design a CA system.

6.1 Gaining overview of the data

Jill opens DELFI and sees the participants’ data broken up by day to allow easy pattern recognition. Figure 4 a shows the pane, called the *Habit View* as analysts can identify repeated contexts. Each separate rectangular segment or “Continuous Context Chunk” represents the time duration for which the user’s context was exactly the same i.e. all labels provided for subsequent sessions were the same (V1). She can easily discern a pattern that most people tend to be “Lying Down” and “Sleeping” at the beginning of the day (00:00 AM) (V2). Since there are large differences in the duration of continuous context, it is difficult to put text on top of some of the smaller chunks as it will be harder to read and may overflow. That is why text is overlaid on chunks that are longer than a certain width. Jill can hover over a chunk to see its context labels (Figure 4 a). She can also identify periods of time where the data was not collected or where the data was present but unlabelled, shown as gray blocks and blue blocks (E2) respectively.

6.2 Finding Mislabelled Data

Jill notices a chunk in the *Habit View* that is bright red (V3). Each chunk has a list of average anomaly for all underlying sessions across all different individual labels comprising the context for that chunk (Figure 3 shows the data attributes associated with individual data session and chunks). The luminance of each chunk is encoded to represent the *highest* anomaly score in that list. Jill hovers over the chunk which then hides all the users data apart from the forty chunks with the highest *feature similarity scores* (V4) (Figure 4 a). Chunks with the exact same context labels are further highlighted by green boxes. This is meant to aid Jill in seeing if the most feature similar chunks are also labelled with the same context. The

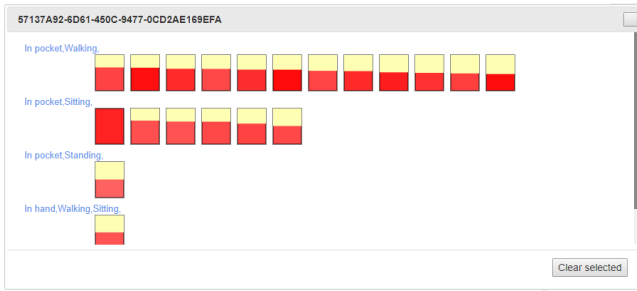


Figure 5: Similarity Dialog: Showing most similar chunks in terms of features.

chunks also have an overlay of color #ffffb3. The longer this overlay is, the *less similar* a chunk is to the hovered over chunk. For example, ■ is more *feature similar* than ■. Jill can see that not all the chunks of similar feature values are labelled as “In Pocket, Sitting”. She further explores this by clicking on the chunk which opens a *Chunk Detail View* (V3, V5) in Figure 4 b. This view separates out the individual labels in a person’s context and the lines inside the bars for the individual labels show the anomaly scores for the time-ordered individual sessions comprising the chunk (V3). The two thinner gray bars on the top show changes in battery charging status and app usage respectively. App usage shows whether there was an app running in the phone foreground when it collected data and battery charging status shows if it was plugged into a USB or an AC electric outlet (V5). Positive and negative instances for app usage and battery charging status are denoted by ■ and ■ respectively. The view also encodes the label providing mechanisms, namely History ■, Active ■ and Notifications ■ with three colors selected in ColorBrewer [11] to ensure that the analyst can easily discern between them. Jill can see that all the individual sessions in this chunk were provided using the “History” option meaning they were labelled retroactively. She notices that while the anomaly scores for the sessions in this chunk for both labels are high, they are particularly higher for “Sitting”. She clicks on the “Most Similar Chunks” button in the Chunk Detail View and she views the *Similarity Dialogue* (Figure 5). This shows her a list of the most similar contexts ordered in terms of the number of occurrence in the top forty most similar chunks (V4). She sees that a dozen of the most similar contexts were labelled as “In Pocket, Walking”. Further, Jill notices that there was a transition in this duration from no app running in the phone foreground to app usage for most of the duration of the context (V5). The break ■ in app usage denotes changing apps open in foreground. When she hovers over the first app in the app usage bar, she sees that it is for the “Maps” app. She is now more confident in her assessment that she has found mislabelled data sessions for at least “Sitting” (E1). To make a note of this discovery, she drags the mouse over the sessions in the context duration across the “Sitting” bar that she thinks is mislabelled and relabels them as “Unlabelled” in the *Relabel Dialog* (Figure 4 c).

6.3 Finding Missing Labels

Jill also wants to find data that has missing labels (E2). She scrolls to another user’s data and notices an unlabelled chunk from morning well into the afternoon. Hovering over the chunk tells her that it is most feature similar to other chunks with “In pocket, Sitting” (V4). She is hesitant to label all the chunks in this long duration with the exact same context. She clicks on the chunk to view it in the *Chunk Detail View* (Figure 6). This shows probability values for three labels that had the highest average overall probability across all the data sessions (see Figure 3 for data description). Jill notices that at the beginning of this chunk (Figure 6 a), the probabilities for the session being “In pocket” were low and jump up as the phone is no longer

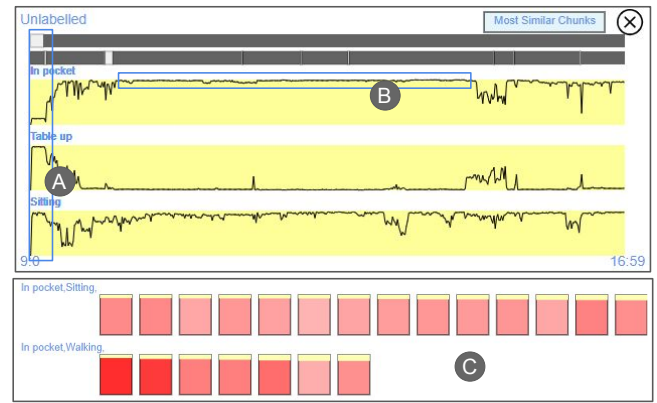


Figure 6: Showing the Chunk Detail view and the Similarity Dialog for an unlabelled chunk of data. The lines in the yellow bars for each label show the probability of that label for the individual sessions comprising the unlabelled chunk.

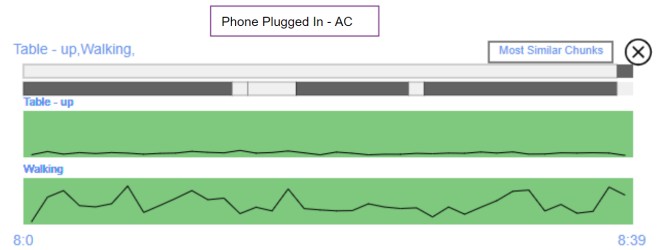


Figure 7: Large variation in the anomaly scores for the two different in the context.

charging (V5). There are also some dips with app usage but for a large portion of the time (Figure 6 b), the probabilities of the session being “In pocket” are quite high. She opens the *Similarity Dialogue* (Figure 6 c) and notices that the most feature similar chunks are those for “In pocket” with “Walking” and “Sitting”, which makes her unsure about the sitting label (V4). Jill notices that there are some dips in the probabilities for chunks later in the chunk. Therefore, Jill selects the portion for “In pocket” for the highest probability sessions (Figure 6 b).

6.4 Finding Large differences in anomaly scores

While exploring these chunks, Jill notices a smaller chunk with a high anomaly score (V3). She hovers over it to discover that it is labelled “Table - up, Walking”. She clicks on it to see that the anomaly scores for data sessions with “Walking” label are consistently more anomalous than those for “Table up” (V3)(Figure 7). She also noticed that for most of the duration, the phone was plugged into an AC connection (V5) which seems likely for the “Table up” label. This leads her to believe that she cannot be certain about the “Walking” label (E1), while not discarding the other label in this context.

Such fine grain analysis shows that visual analytics can fill in the pitfalls of relying solely on automated anomaly measure calculation approaches.

7 EXPERT EVALUATION

To evaluate DELFI, we invited three experts in the field of human context data modelling and inferring. The experts were given a

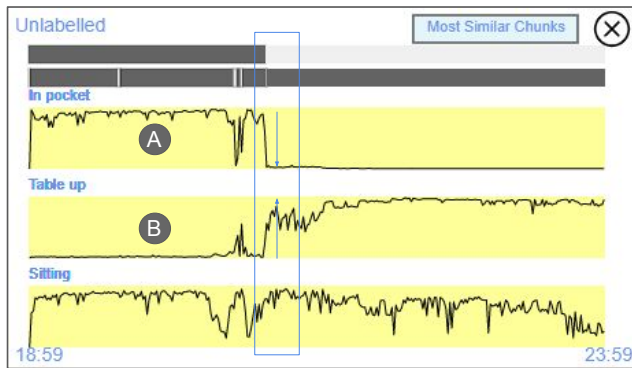


Figure 8: There is a major spike in the probability values for “Table - up” and a major drop in the probability for “In pocket”.

tutorial for using DELFI. They were then guided to view the chunks that Jill had viewed to find mislabels. The experts were asked to verbalize their thought process as they explored the data and not to constrain themselves to the use cases that we provide but rather to expand on them and explore other data sessions as well. After a tutorial, they were able to note a clear trend for when people labelled “Sleeping”. As they went through the use cases, they generally came to the same conclusions as Jill about the types of mislabelling that occurred.

They liked the multi-faceted approach of showing anomaly scores in conjunction with a measure of feature similarity along with the ability to flexibly select the session for relabelling and unlabelling. One expert noted how a naive version of feature similarity linking would be to just cluster data but its high dimensionality makes it difficult which is why showing the data in a multi-pane view allows them to gain a better picture. During their exploration of the data, they found an unlabelled chunk (Figure 8) where there were rapid changes in the probability values for “In pocket” (steep decline Figure 8 a) and “Table up” (shortened from “Phone on Table - Facing up”) (steep rise Figure 8 b). The experts were quick to note that this happened around the same time that the phone was plugged in. This lead them to conclude that this sharp change connotes a transition between “In pocket” to “Table up”.

One expert who specializes in transitions in human context suggested using a method called “Change Point Detection” to find such instances of transition and highlighting them visually. Overall, the experts liked the approach and indicated interest in following future work.

8 DISCUSSION AND LIMITATIONS

We have shown the utility of VA in analyzing human context data. However, there are some limitations. This approach is time consuming and does not allow for bulk selection of data. Future implementations may incorporate some top level filtering mechanisms such as excluding all data with certain co-occurring labels, highlighting only the data with some app usage etc. However, a limitation to that is people have specific phone using habits and typically individualized modelling of data proves to be more reliable than all inclusive modelling [17, 21].

Additionally, a limitation of our similarity linking method occurs when a given chunk has been mislabeled, and in reality two or more distinct contexts occurred within that chunk. Our feature similarity linking method would find the chunks that are most similar to that given mislabeled chunk as a whole, as we do not have a way of distinguishing that there are multiple distinct contexts within that chunk. Change point detection techniques are known to be able to segment individual contexts in a continuous stream of activity

data [1]. In future work, a change point detection method can be implemented to identify when there are multiple contexts within one labeled continuous context chunk.

9 CONCLUSION

We motivated the paper by describing the limitations of prior human behavior visualization systems for dealing with user-labeled ground truth data. To overcome these limitations, we introduced DELFI, an interactive tool designed to explore and find instances of mislabeled human context data. To that end, we developed a framework that identifies mislabeled instances and determines the likely true labels using an anomaly score and a novel multi-feature similarity score in conjunction. We proposed a visual metaphor Multi-Feature Similarity Linking to link data in terms of feature similarity to help analysts find data with mislabeled ground truth. DELFI’s utility was illustrated by walking through multiple use cases and with expert feedback.

ACKNOWLEDGMENTS

The work presented in this paper was funded by DARPA WASH program HR00117S0032

REFERENCES

- [1] S. Aminikhanghahi, T. Wang, and D. J. Cook. Real-time change point detection with application to smart home time series data. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):1010–1023, 2018. doi: 10.1109/TKDE.2018.2850347
- [2] D. Archambault, D. Greene, P. Cunningham, and N. Hurley. Theme-crowds: Multiresolution summaries of twitter usage. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pp. 77–84. ACM, 2011. doi: 10.1145/2065023.2065041
- [3] S. Bae, T. Chung, D. Ferreira, A. K. Dey, and B. Suffoletto. Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. *Addictive behaviors*, 83:42–47, 2018. doi: 10.1016/j.addbeh.2017.11.039
- [4] J. Bekker and J. Davis. Learning from positive and unlabeled data: A survey. *arXiv preprint arXiv:1811.04820*, 2018.
- [5] N. Cao, C. Shi, S. Lin, J. Lu, Y.-R. Lin, and C.-Y. Lin. Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *IEEE transactions on visualization and computer graphics*, 22(1):280–289, 2015. doi: 10.1109/TVCG.2015.2467196
- [6] Y.-J. Chang, G. Paruthi, H.-Y. Wu, H.-Y. Lin, and M. W. Newman. An investigation of using mobile and situated crowdsourcing to collect annotated travel activity data in real-world settings. *International Journal of Human-Computer Studies*, 102:81–102, 2017. doi: 10.1016/j.ijhcs.2016.11.001
- [7] S. Chen, L. Lin, and X. Yuan. Social media visual analytics. In *Computer Graphics Forum*, vol. 36, pp. 563–587. Wiley Online Library, 2017. doi: 10.1111/cgf.13211
- [8] D. D. Ebert, P. Cuijpers, R. F. Muñoz, and H. Baumeister. Prevention of mental health disorders using internet-and mobile-based interventions: a narrative review and recommendations for future research. *Frontiers in psychiatry*, 8:116, 2017. doi: 10.3389/fpsy.2017.00116
- [9] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220. ACM, 2008. doi: 10.1145/1401890.1401920
- [10] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001. doi: 10.1214/aos/1013203451
- [11] M. Harrower and C. A. Brewer. Colorbrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. doi: 10.1179/000870403235002042
- [12] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Tenth International Conference on Information Visualisation (IV’06)*, pp. 9–16. IEEE, 2006. doi: 10.1109/IV.2006.31

- [13] K. Kucher, C. Paradis, and A. Kerren. The state of the art in sentiment visualization. In *Computer Graphics Forum*, vol. 37, pp. 71–96. Wiley Online Library, 2018. doi: 10.1111/cgf.13217
- [14] C. Lipizzi, L. Iandoli, and J. E. R. Marquez. Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers reactions to the launch of new products using twitter streams. *International Journal of Information Management*, 35(4):490–503, 2015. doi: 10.1016/j.ijinfomgt.2015.04.001
- [15] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008. doi: 10.1109/icdm.2008.17
- [16] S. Liu, C. Chen, Y. Lu, F. Ouyang, and B. Wang. An interactive method to improve crowdsourced annotations. *IEEE Transactions on Visualization Computer Graphics*, 25(01):235–245, jan 2019. doi: 10.1109/TVCG.2018.2864843
- [17] A. Mehrotra, M. Musolesi, R. Hendley, and V. Pejovic. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 813–824. ACM, 2015. doi: 10.1145/2750858.2807544
- [18] P. H. Nguyen, C. Turkay, G. Andrienko, N. Andrienko, O. Thonnard, and J. Zouaoui. Understanding user behaviour through action sequences: From the usual to the unusual. *IEEE Transactions on Visualization and Computer Graphics*, 25(9):2838–2852, sep 2019. doi: 10.1109/tvcg.2018.2859969
- [19] D. Nielsen. Tree boosting with xgboost-why does xgboost win” every” machine learning competition? Master’s thesis, NTNU, Norway, 2016.
- [20] P. J. Polack Jr, S.-T. Chen, M. Kahng, K. D. Barbaro, R. Basole, M. Sharmin, and D. H. Chau. Chronodes: Interactive multifocus exploration of event sequences. *ACM Transactions on Interactive Intelligent Systems*, 8(1):1–21, 2018. doi: 10.1145/3152888
- [21] X. Su, H. Tong, and P. Ji. Activity recognition with smartphone sensors. *Tsinghua science and technology*, 19(3):235–249, June 2014. doi: 10.1109/TST.2014.6838194
- [22] Y. Vaizman, K. Ellis, and G. Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE Pervasive Computing*, 16(4):62–74, October 2017. doi: 10.1109/MPRV.2017.3971131
- [23] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel. Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, pp. 554:1–554:12, 2018. doi: 10.1145/3173574.3174128
- [24] Y. Vaizman, N. Weibel, and G. Lanckriet. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):168, 2018.
- [25] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp ’14 Adjunct*, pp. 3–14. ACM, 2014. doi: 10.1145/2632048.2632054
- [26] 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, Nov. 2009. doi: 10.1109/TVCG.2009.187
- [27] R. Younes, M. Jones, and T. L. Martin. Classifier for activities with variations. *Sensors*, 18(10), Oct 2018. doi: 10.3390/s18103529
- [28] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins. #fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE transactions on visualization and computer graphics*, 20(12):1773–1782, Dec 2014. doi: 10.1109/TVCG.2014.2346922