

Literature Review

Recognition of Emotions Based on Sensor Data

Group G01

29. November 2022

We are conducting a systematic and reproducible literature review in order to get a better understanding of the topics surrounding our research project on Sensor Based Emotion Recognition. We are basing our methodology on a framework by Templier and Paré [1]. We will begin by identifying key phrases. We will then perform a broad search using these key phrases. Next, we will assess the relevancy of the works we found and select a few to be analysed thoroughly. We will be assessing the quality of those works and extract relevant information.

1 Formulating the problem

Our goal is to get a better understanding of the topics surrounding our research project. Our work is based on a project by our supervisor [2]. We will use similar methods, but apply them to the ExtraSensory Dataset [3]. We want to find similar or otherwise relevant works on emotion recognition, works using similar machine learning approaches, works dealing with sensor data from wearable sensors and works that have used the ExtraSensory Dataset. We have identified several keyphrases we will pay special attention to in our initial search.

Keyphrases:

- Emotion Recognition
- Emotion Detection
- Wearable Sensors
- ExtraSensory Dataset
- Human Activity Recognition
- Machine Learning
- Clustering
- Classification
- Supervised Learning

2 Searching the literature

Using the keyphrases mentioned above, we have performed a broad search for related works and listed our findings below, in addition to the base-paper, i.e. the paper by Popko [2], and other works referenced there.

| Authors | Title | Source | Notes |
|---|---|---|---|
| M. Popko, S. Bader, S. Lüdtke, T. Kirste | Discovering behavioral predisposition in data to improve human activity recognition | given by the supervisor | Starting point for our research |
| Y. Vaizman, Ellis, K., and Lanckriet, G | Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches | https://ieeexplore.ieee.org/document/8090454 | Original paper of the provided data source |
| D. Blei, A. Ng, and M. Jordan | Latent Dirichlet Allocation | https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf | Referenced by supervisor |
| F. Cruciani, C. Sun, S. Zhang, et al. | A Public Domain Dataset for Human Activity Recognition in Free-Living Conditions | https://ieeexplore.ieee.org/abstract/document/9060182/authors | Found by searching "ExtraSensory Dataset" on Google Scholar |
| R. Alam, A. Bankole, M. Anderson, and J. Lach. | Multiple-Instance Learning for Sparse Behavior Modeling from Wearables: Toward Dementia-Related Agitation Prediction | https://ieeexplore.ieee.org/document/8856502 | Referenced by supervisor |
| F. Cruciani, A. Vafeiadis, C. Nugent et al. | Feature learning for Human Activity Recognition using Convolutional Neural Networks | https://link.springer.com/article/10.1007/s42486-020-00026- | Found by searching "ExtraSensory Dataset" on Google Scholar |
| A. Dzedzickis, A. Kalkauskas, V. Bucinskas | Human Emotion Recognition: Review of Sensors and Methods | https://www.mdpi.com/1424-8220/20/3/592 | Found by searching "emotion recognition sensors" on Google Scholar |
| M. Sultana, M. Al-Jefri, J. Lee | Using Machine Learning and Smartphone and Smartwatch Data to Detect Emotional States and Transitions: Exploratory Study | https://mhealth.jmir.org/2020/9/e17818/ | Found by searching "ExtraSensory Dataset" on Google Scholar |
| S. Lüdtke, F. Rueda, W. Ahmed, G. Fink, and T. Kirste | Human Activity Recognition using Attribute-Based Neural Networks and Context Information | https://arxiv.org/abs/2111.04564 | Referenced by supervisor |
| M. Z. Rodriguez, C. H. Comin, D. Casanova et al. | Clustering algorithms: A comparative approach | https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0210236 | Found by searching "clustering algorithms" on Google Scholar |
| A. Lengyel, Z. Botta-Dukát | Silhouette width using generalized mean—A flexible method for assessing clustering efficiency | https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/ece3.5774 | Found by searching 'assessing Clustering methods' |
| C. Schaffer | Selecting a Classification Method by Cross-Validation | https://link.springer.com/article/10.1007/BF00993106 | Found by searching: 'clustering methods' |
| A. M. Khan, A. Tufail, A. M. Khattak and T. H. Laine | Activity Recognition on Smartphones via Sensor-Fusion and KDA-Based SVMs | https://journals.sagepub.com/doi/10.1155/2014/503291 | Found by using "Litmaps" with the seed article [3] |
| D. Shi, Xi Chen, J. Wei, R. Yang | User Emotion Recognition Based on Multi-Class Sensors of Smartphone | https://ieeexplore.ieee.org/abstract/document/7463770 | Found by searching "emotion recognition smartphone" on Google Scholar |
| Wikipedia contributors | Cluster analysis | https://en.wikipedia.org/wiki/Cluster_analysis#Evaluation_and_assessment | After searching for "Silhouette width" because of the paper by Lengyel et al., a scikit-learn user guide about clustering ¹ was found. It referred to the Rand index article on Wikipedia, and it, in turn, referred to this site. |

¹<https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation>

3 Screening for inclusion

Using the list of Literature we identified above, we perform a more thorough screening and select a few works to be analysed in the next section.

| Ref | Authors | Title | Is it relevant for further analysis? |
|-----|---|---|---|
| [2] | M. Popko, S. Bader, S. lüdtke, T. Kirste | Discovering behavioral predisposition in data to improve human activity recognition | Yes, as it is the basis for our work. |
| [3] | Y. Vaizman, Ellis, K., and Lanckriet, G | Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches | Yes. This paper goes into more into depth on the technical aspect. As we do not focus on developing and implementing the electronics to gather the data, it is not of much interest. |
| | D. Blei, A. Ng, and M. Jordan | Latent Dirichlet Allocation | No, because this paper goes too much into depth of LDA, but the method itself is relevant for us. |
| [4] | F. Cruciani, C. Sun, S. Zhang, et al. | A Public Domain Dataset for Human Activity Recognition in Free-Living Conditions | Yes. They are also working with the ExtraSensory dataset, but using convolutional neural networks for classifying. Although there are some interesting points how they optimize and validate their methods. |
| [5] | R. Alam, A. Bankole, M. Anderson, and J. Lach. | Multiple-Instance Learning for Sparse Behavior Modeling from Wearables: Toward Dementia-Related Agitation Prediction | Yes. They are using Multiple instance learning on motion sensor data from a wristband to identify episodes of agitation. They are using a different category of machine learning algorithms than our work and are not trying to identify emotion in general, but this paper is still somewhat relevant to us. |
| | F. Cruciani, A. Vafeiadis, C. Nugent et al. | Feature learning for Human Activity Recognition using Convolutional Neural Networks | No. Although it also uses the ExtraSensory dataset (here, mainly the audio and activity part), it focuses too much on convolutional neural networks and its performance, that will not be part of our research project. Other relevant information can also be retrieved from the website of the dataset. A more compelling source that also used our dataset would be Sultana et al. |
| [6] | A. Dziedzickis, A. Kaklauskas, V. Bucinskas | Human Emotion Recognition: Review of Sensors and Methods | No really. It has some interesting ideas and methods that are closely related to our reference paper. But for our specific project the ideas and methods presented are not of too much interest. |
| [7] | M. Sultana, M. Al-Jefri, J. Lee | Using Machine Learning and Smartphone and Smartwatch Data to Detect Emotional States and Transitions: Exploratory Study | Yes! This paper is very interesting because the same dataset is used and some machine learning algorithms are tested on it. Therefore it is a very important paper. |
| | S. Lüdtke, F. Rueda, W. Ahmed, G. Fink, and T. Kirste | Human Activity Recognition using Attribute-Based Neural Networks and Context Information | No. This study focuses too much on Neural Networks and furthermore uses a different dataset. Also its recognizing human activities and not emotions. The only interesting thing for us might be, that they showed how context information can lead to a better results. |
| | M. Z. Rodriguez, C. H. Comin, D. Casanova et al. | Clustering algorithms: A comparative approach | No. They compare and evaluate 9 different clustering algorithms. Clustering is an important part of the pipeline we want to implement. When it comes to choosing different clustering approaches and evaluating them, this is a paper, we can fall back on. |
| | A. Lengyel, Z. Botta-Dukát | Silhouette width using generalized mean—A flexible method for assessing clustering efficiency | No, but it featured an evaluation metric that we can use for our project. But everything is not very relevant (comparisons of different means for using in the Silhouette width metric). More evaluation metrics can be found in the Wikipedia article about "Cluster analysis" |
| | C. Schaffer | Selecting a Classification Method by Cross-Validation | No. This paper proposes a few interesting ideas in regards to selecting a good classification method. Those are might have a few implications, but this paper seems to be a little bit outdated. |
| | A. M. Khan, A. Tufail, A. M. Khattak and T. H. Laine | Activity Recognition on Smartphones via Sensor-Fusion and KDA-Based SVMs | No. This study tries to implement a system which recognizes human activities for smartphones. But its not quite relevant for us, because they are not using clustering and a main part of the paper focuses on extracting features. |
| [8] | D. Shi, Xi Chen, J. Wei, R. Yang | User Emotion Recognition Based on Multi-Class Sensors of Smartphone | Yes, because the problem they tackle is very similar to our own. They are using classification to map sensor data gathered from smartphones to emotion. |
| [9] | Wikipedia contributors | Cluster analysis | Yes, because it has a comprehensive overview about the different clustering evaluation methods. The decision to use this Wikipedia article is that it more practical to use than Lengyel et al. |

4 Extracting data

In this section we are extracting relevant information from the papers we deemed most relevant for our project in the section above.

4.1 Discovering behavioral predisposition in data to improve human activity recognition [2]

- For many domains, human activity recognition (HAR) is crucial.
- With the help of wearable sensors, HAR can be used to assess symptoms of patients with dementia.
- However, as the success with sensor data on its own is limited, the accuracy needs to be improved through the use of external knowledge.
- For example, daily or weekly repeating behaviours called behavioral predispositions (BPDs) are used to increase HAR performance substantially.
- The **pipeline**:
 - The calculated distribution of annotations (histograms) are clustered to BPDs.
 - After that, a classifier is trained for each of them.
 - Parts of the motion sensor data are selected, and these features are then used by the classifier to predict the behaviour.
- Clustering can be used to identify and group times where a person has had similar behaviours globally (over all days).
- Now that we have the BPDs, and thus a bit of prior information on possible behaviour, a classifier can be trained for each of them.
- So, motion sensor data is mapped to behaviours. By training a classifier on a certain BDP instead on the entire dataset, we can make use of a person's tendency to behave similarly in a certain kind of situation. By that, the accuracy of the prediction is improved.
- This paper also contains ideas that could be relevant (for us), too
 - The **Jensen-Shannon divergence** could be a better alternative to the **k-means** clustering algorithm
 - The Latent Dirichlet Allocation (LDA) can be used
 - Classifiers that use the distributions over BPDs have to be considered or mixed-effects neural networks
 - For practical use, the right k for clustering has to be found in order to keep the classification error low
 - Hidden-Markov models were also named
 - Decision trees and convolutional neural networks were used in another work to "weight the predicted state based on the sensor data"
 - External context information improves HAR performance, it is even enough when the day is segmented in different periods of time

4.2 Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches

- **Capture behavior using everyday devices:** They try to use common place items (phones and smartwatches) to built-in sensory data.
- **Understanding conditons:**
 - **Naturally used devices:** Introducing a foreign device would put a burden on the user and might affect their natural behaviour, thus naturally used devices are preferable
 - **Unconstrained device placement:** The placement of said devices can improve recognition success. But practically speaking, only natural placements are to be considered.
 - **Natural evironment:** In this context we ought to only record their behavior for their own schedule.
 - **Natural behavioral content:** Here, we look into subjects, that are given stripted tasks to simulate natural behavior.
- **The Context Recognition System:**
 - System is based on five different measurements namely, accelerometer, gyroscope, location, audio, phone-state sensors and accelerometer measurements.
 - For any given minute the system sampled measurements for each respective sensor.
 - Given these inputs the task is to detect a coherent context label.
 - every minute was sampled independently and every model was labeled separately .
 - Only simple methods were implemented to demonstrate the potential of context recognition.

- **Single-sensor Classifiers:**

- These classifiers were used to help understand sensor-specific features.
- While also looking at them independently and how informative they are for a given context label.
- A linear classifier was used that outputs binary continuous values which are interpreted as probability.

- **Sensor Fusion:**

- early fusion classifier is used, which combines the information from multiple sensors prior to classification.
- combines probability to avoid the influence from irrelevant sensors.

4.3 A Public Domain Dataset for Human Activity Recognition in Free-Living Conditions

- This study used the Extrasensory dataset for training and validating a convolutional neural network to detect human activities.
- For **validation** they used a 5-fold cross validation. To overcome overfitting they stopped the training, when the accuracy began to decrease on the validation set
- To optimize their results they were using a Stochastic Gradient Descent, "which has been observed to provide better generalization on unseen data"
- They used a new dataset, which they have created, to test their classifier.

4.4 Multiple-Instance Learning for Sparse Behavior Modeling from Wearables: Toward Dementia-Related Agitation Prediction

- They are using Multiple instance learning on motion sensor data from a wristband, to identify episodes of agitation in dementia patients.
- **Extracting Features from Motion Sensor Data:** They used a window width of 60-seconds with 50% overlap to acquire a three-dimensional signal window every 30 seconds. They used a median filter to remove the speckle noise and then applied a bandpass FIR filter to reduce motion artifacts. They extracted mean, median, max, standard deviation, variance, rms level, and interquartile range.
- **Evaluation:** They hold-out 30% of their data as test-data and use 5-fold cross-validation to train on the remaining 70%. They use accuracy, F-score and AUC to compare the performance of their models. They compare three Multiple instance learning models (APR, MI-SVM, and MIL-Boost) and three single instance learning models (SVM, AdaBoost, random forest)

4.5 Using Machine Learning and Smartphone and Smartwatch Data to Detect Emotional States and Transitions: Exploratory Study

- This exploratory study searched for the relationship between everyday context and emotional transitions.
- Moreover, **this study uses our ExtraSensory Dataset** to achieve this.
- This dataset includes 49 emotions reported by 18 persons (only these subjects reported more than 1000 samples and their proportion of missing data is under 90%)
- They included all signals by the devices
- They built personalized and general models
- They mapped the 49 emotions to 3 emotion dimensions, so an emotion can be either strong or low in one dimension (e.g, discordant or pleased) - so they have 6 emotional states (annotation: could mean $k = 6$ for the clustering algorithm, because the emotions are grouped in these states) - so they did not use any clustering algorithm
- Additionally, they wanted to solve the transition detection problem (that is "binary classification"), which will not be relevant to us. But they defined the state detection as a "**multiclass classification problem**", that is nearly equal to our problem
- They used different algorithms and shared their results on them:
 - These 5 **supervised machine learning algorithms** were used: logistic regression, random forest, XG-Boost, CatBoost and multilayer perceptron.
 - For general models: six-fold, leave-3-people-out cross-validation (hyperparameters were tuned by the F_1 score)

- For personalized models: five-fold, stratified cross-validation
- For state detection: The area under the receiver operating characteristic (AUROC) curve is far worse for general models (60,55%, with logistic regression) then for personalized models in average (96,33%, with CatBoost)
- All ML algorithms for personalized models performed somehow similar (AUROC varies from 93,74% to 96,33%)
- The **most important features** to get better results are: spatiotemporal context, phone state, motion-related information (Annotation: Phone state data was already assumed by Jonny to be important)
- Sidenote: Lifestyle impacts how predictable the emotions are
- As there were imbalances in the dataset, two oversampling methods were used to decrease the imbalance.
- In the study, different features were used:
 - **138 motion features** were calculated using the accelerometer, gyroscope, magnetometer by the smart-phone and the accelerometer and compass by the smartwatch
 - **Audio** data was also used
 - **17 location features**, and 3 additional features were calculated (Annotation: The latter seem not relevant for us, as they seem to be used for emotional transition detection)
 - **Environmental data** had many missing values
 - They calculated 5 more **temporal** features with the *timestamp* feature (e.g, hour of the day). They sampled all data to time intervals of 5 minutes.
 - Trusting the subjects, the 51 **contextual features** were used (e.g, EATING)
- In an addition to this study, they released a "multimedia appendix" that includes, among others, the **list of features** (Table 2), **hyper-parameter search grids** (Table 4), **Results for the ML methods** (general: Table 6, personalized: Table 8). And, a publicly available GitHub repository² contains the code used for this study.

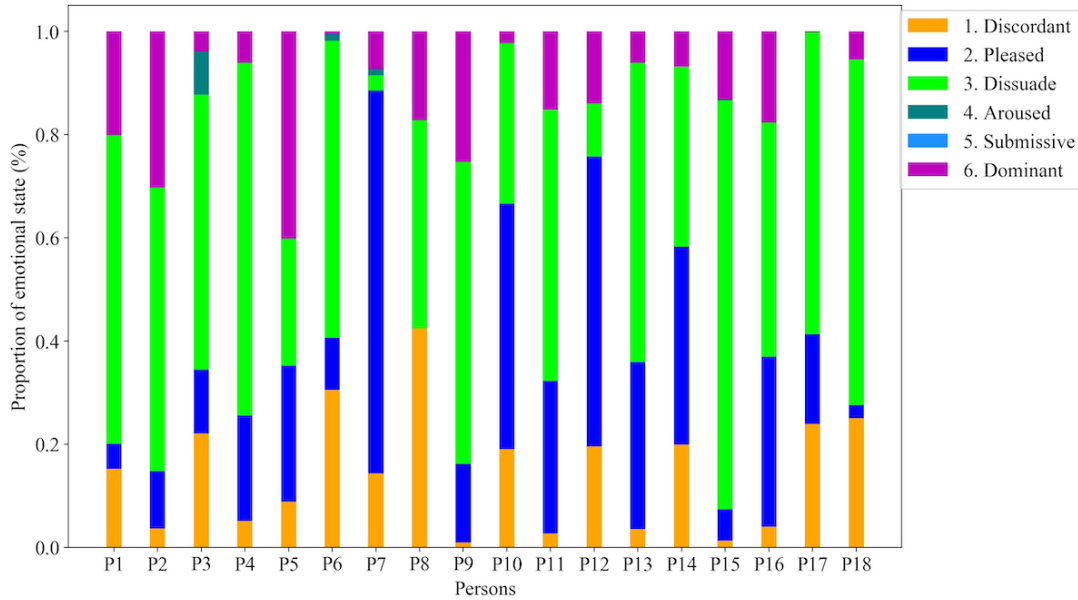


Figure 1: Proportion of the emotional states [7]

4.6 Human Emotion Recognition: Review of Sensors and Methods

• Emotions Evaluation Methods

- techniques used for emotion recognition:
 - * self-report techniques based on self-assessment
 - * machine assessment based on measurement of various parameters of the human body
- **Signal Analysis and Features Extraction Methods**

²https://github.com/data-intelligence-for-health-lab/mHealth_Emootional_State-Transition_Detection_Public

Table 14. Analysis of previous studies on emotion recognition.

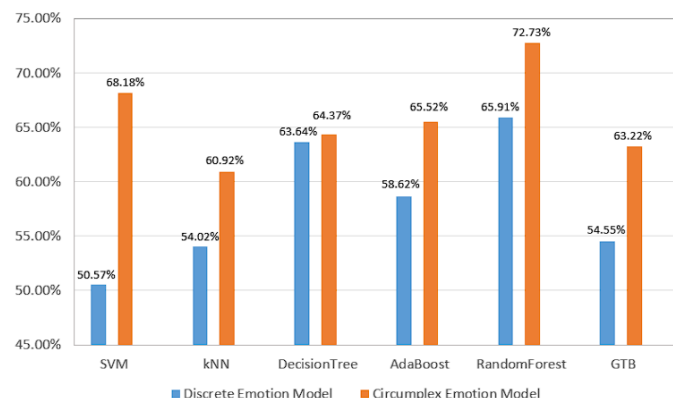
| Emotions | Measurement Methods | Data Analysis Methods | Accuracy | Ref. |
|---|---|--|--|-------|
| Sadness, anger, stress, surprise | ECG, SKT, GSR | SVM | Correct-classification ratios were 78.4% and 61.8%, for the recognition of three and four categories, respectively | [133] |
| Sadness, anger, fear, surprise, frustration, and amusement | GSR, HRV, SKT | KNN, DFA, MBP | KNN, DFA, and MBP, could categorize emotions with 72.3%, 75.0%, and 84.1% accuracy, respectively | [184] |
| Three levels of driver stress | ECG, EOG, GSR and respiration | Fisher projection matrix and a linear discriminant | Three levels of driver stress with an accuracy of over 97% | [126] |
| Fear, neutral, joy | ECG, SKT, GSR, respiration | Canonical correlation analysis | Correct-classification ratio is 85.3%. The classification rates for fear, neutral, joy were 76%, 94%, 84% respectively | [185] |
| The emotional classes identified are high stress, low stress, disappointment, and euphoria | Facial EOG, ECG, GSR, respiration, | SVM and adaptive neuro-fuzzy inference system (ANFIS) | The overall classification rates achieved by using tenfold cross validation are 79.3% and 76.7% for the SVM and the ANFIS, respectively. | [122] |
| Fatigue caused by driving for extended hours | HRV | Neural network | The neural network gave an accuracy of 90% | [186] |
| Boredom, pain, surprise | GSR, ECG, HRV, SKT | Machine learning algorithms: linear discriminate analysis (LDA), classification and regression tree (CART), self-organizing map (SOM), and SVM | Accuracy rate of LDA was 78.6%, 93.3% in CART, and SOMs provided accuracy of 70.4%. Finally, the result of emotion classification using SVM showed accuracy rate of 100.0%. | [187] |
| The arousal classes were calm, medium aroused, and activated and the valence classes were unpleasant, neutral, and pleasant | ECG, pupillary response, gaze distance | Support vector machine | The best classification accuracies of 68.5 percent for three labels of valence and 76.4 percent for three labels of arousal | [188] |
| Sadness, fear, pleasure | ECG, GSR, blood volume pulse, pulse. | Support vector regression | Recognition rate up to 89.2% | [189] |
| Frustration, satisfaction, engagement, challenge | EEG, GSR, ECG | Fuzzy logic | 84.18% for frustration, 76.83% for satisfaction, 97% for engagement, 97.99% for challenge | [190] |
| Terrible, love, hate, sentimental, lovely, happy, fun, shock, cheerful, depressing, exciting, melancholy, mellow | EEG, GSR, blood volume pressure, respiration pattern, SKT, EMG, EOG | Support Vector Machine, Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN) and Meta-multiclass (MMC), | The average accuracies are 81.45%, 74.37%, 57.74% and 75.94% for SVM, MLP, KNN and MMC classifiers respectively. The best accuracy is for 'Depressing' with 85.46% using SVM. Accuracy of 85% with 13 emotions | [191] |

4.7 User Emotion Recognition Based on Multi-Class Sensors of Smartphone

- They gathered sensor data from smartphones and used classification to identify emotion.
- **Defining a model for Emotion:** Emotion can be defined as a discrete number or continuous values of emotions. Using the Circumplex Emotion Model, every emotion can be classified by sole measurement of pleasure and level of activity. They were both tested and showed promising results using the Circumplex model.
- **Missing Data Handling:** They recorded data from 12 People, over 30 days, between the times of 7:30 and 22:30. Out of 1080 segments of 5 hour length, they selected 743 to be used and discarded the rest. They did not specify their exact screening process.

- **Extracting Features from Motion Sensor Data:** They segmented their data into 50% overlapping sliding windows of 5 second length. They extracted 9 features from each sensor in each window:
 - maximum
 - minimum
 - mean
 - standard deviation
 - wave number
 - peak mean
 - trough mean
 - the maximum difference between the peak and trough
 - the minimum difference between the peak and trough
- **Evaluation of Classification Approaches:** To evaluate their results they used 5-Fold Cross-validation. They tested 6 different algorithms and applied each on both the discrete and Circumplex Emotion Model:
 - Support Vector Machine (SVM)
 - k-Nearest Neighbor (kNN)
 - Decision Tree
 - AdaBoost
 - Random Forest
 - Gradient Tree Boosting (GTB)

As seen in their figure shown below, the Circumplex Model results in higher accuracy across the board, while the random forest approach resulted in the highest accuracy for both models.



4.8 "Clustering analysis" on Wikipedia [9]

- The focus here is on the "Evaluation and assessment" chapter.
- Some general findings:
 - Clustering evaluation is as hard as clustering itself
 - There is a difference between internal (single quality score) and external evaluation (comparison to already available ground truth classification)
 - Annotation: manual and indirect evaluation seem to be irrelevant for us, but we have to interpret these scores (making these evaluations subjective)
 - Evaluation has one big problem: In order to calculate these scores, clustering has to be done. So these scores should rather be used for comparing the similarity of the optimization problems
- How to use **internal evaluation**:
 - These scores are high when the clusters have high similarity within a cluster and low similarity between clusters
 - We have to always keep these problems in mind:
 - * If a program has a high score, it does not always mean that it is effective for information retrieval
 - * Evaluation metrics can be biased to algorithms with similar cluster models

- * Higher scores \neq more valid results
- * Many metrics assume convex clusters
- These indices can be considered:
 - * **Davies-Bouldin index:** The smaller the index of each cluster, the better (calculates distances)
 - * **Dunn index:** The higher the intra-cluster similarity and the lower the inter-cluster similarity, the higher is the Dunn index
 - * **Silhouette coefficient:** This clustering evaluation metric is very promising for our project, as it can be used to determine the optimal number of clusters (also see the scikit-learn user guide referenced above)
- How to use **external evaluation:**
 - These scores need benchmark sets or labels to work (e.g, known class names)
 - There are several scores available, such as:
 - * **Rand index:** Score on how similar the clusters are to the benchmark. A better alternative would be the
 - * **F-measure:** By weighting the recall with a parameter, the contribution of false negatives can be balanced.
 - * **Fowlkes-Mallows index:** The higher the clusters and the benchmark classifications are, the higher is this score (also strongly recommended for our project)
 - * **Confusion matrix:** At the end, the confusion matrix is a quick visual way to compare the results of the clustering algorithm

5 Analysis and Summary

In this section we will analyze our findings by first identifying relevant tasks with significant overlap between papers. We are then summarizing how the different works proceeded for each task.

5.1 Defining Emotions

Emotions can be defined discretely or continuously. I.e., having distinct number of emotions with no overlap, or determining an emotion based on some continuous metrics. [8] tried to compare both approaches. They used one model defining emotions discretely, and one model where each emotion can be classified solely by measurement of pleasure and level of activity. This continuous model is based on the Circumplex Emotion Model. During evaluation, the Circumplex generally resulted in higher accuracy. [7] mapped the 49 emotions to 3 emotion dimensions, so an emotion can be either strong or low in one dimension (e.g, discordant or pleased) - so they have 6 emotional states.

5.2 Missing Data Handling

[8] recorded data from 12 people, over 30 days, between the times of 7:30 and 22:30. Out of 1080 segments of 5 hour length, they selected 743 to be used and discarded the rest. They did not specify their exact screening process. [7] included only records of subjects that reported more than 1000 samples and whose proportion of missing data is under 90%. For general data handling, they calculated 5 more temporal features with the *timestamp* feature (e.g, hour of the day) and they sampled all data to time intervals of 5 min.

5.3 Evaluation methods

A lot of the papers we inspected, used different methods when evaluating their results. How did they separate their data into training, validation and testing data sets? How many different approaches did they choose to evaluate? E.g. different clustering algorithms, classifiers, or ways of categorizing emotions. Which metrics did they choose to evaluate on?

Separating train/test data: [3] used 5-Fold cross-validation, where each fold had 48 users for training and 12 users for testing. [8] simply used 5-Fold Cross-validation. [7] used 6-Fold Cross-validation, but instead of equally dividing their samples, they always left out samples from 3 of their 18 subjects. [4] used 5-fold Cross-validation on the ExtraSensory Dataset to train their model, but then they tested on a completely new set of data they collected from 10 subjects. Most interestingly, [5] held-out 30% of their data as test-data and then used 5-fold cross-validation to train on the remaining 70%. This approach lets you get a reference for the performance of your model so that you can improve upon it, while avoiding indirectly overfitting your model by still having a separate test set.

What approaches were evaluated: These 5 supervised machine learning algorithms were used by [7]: logistic regression, random forest, XGBoost, CatBoost and multilayer perceptron. [5] compared three Multiple instance learning models (APR, MI-SVM, and MIL-Boost) and three single instance learning models (SVM, AdaBoost, random forest). [8] tested 6 different classification algorithms (Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Decision Tree, AdaBoost, Random Forest, Gradient Tree Boosting (GTB)) and applied each on both their discrete and Circumplex Emotion Model. Using the Circumplex Model, they achieved higher accuracy across the board, while the random forest approach resulted in the highest accuracy for both models.

Evaluation metrics: [5] used accuracy, F-score and AUC to compare the performance of their models, while [8] just focused on accuracy. [9] listed even more metrics. For internal evaluation (without the use of the ground truth), the Davies-Bouldin index, the Dunn index and the (most relevant) Silhouette coefficient should be considered. As we have access to the ground truth of the test data, the external evaluation metrics Rand index, F-measure and Fowlkes-Mallows index are important for us. A great visualisation method is also proposed: the confusion matrix.

References

- [1] M. Templier and G. Paré, “A framework for guiding and evaluating literature reviews,” *Commun. Assoc. Inf. Syst.*, vol. 37, 2015.
- [2] M. Popko, S. Bader, S. Lüdtke, and T. Kirste, “Discovering behavioral predispositions in data to improve human activity recognition,” 2022.
- [3] Y. Vaizman, K. Ellis, and G. Lanckriet, “Recognizing detailed human context in the wild from smartphones and smartwatches,” *IEEE Pervasive Computing*, vol. 16, no. 4, pp. 62–74, 2017.
- [4] F. Cruciani, C. Sun, S. Zhang, C. Nugent, C. Li, S. Song, C. Cheng, I. Cleland, and P. Mccullagh, “A public domain dataset for human activity recognition in free-living conditions,” in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*. IEEE, 2019.
- [5] R. Alam, A. Bankole, M. Anderson, and J. Lach, “Multiple-instance learning for sparse behavior modeling from wearables: Toward dementia-related agitation prediction,” *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2019, pp. 1330–1333, 2019.
- [6] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, “Human emotion recognition: Review of sensors and methods,” *Sensors (Basel)*, vol. 20, no. 3, p. 592, 2020.
- [7] M. Sultana, M. Al-Jefri, and J. Lee, “Using machine learning and smartphone and smartwatch data to detect emotional states and transitions: Exploratory study,” *JMIR MHealth UHealth*, vol. 8, no. 9, p. e17818, 2020.
- [8] D. Shi, X. Chen, J. Wei, and R. Yang, “User emotion recognition based on multi-class sensors of smartphone,” in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*. IEEE, 2015.
- [9] Wikipedia contributors, “Cluster analysis — Wikipedia, the free encyclopedia,” https://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=1116924542, 2022, [Online; accessed 25-November-2022].

Reproducibility annotation by us: The bibliography was generated with bibguru.com.