Predicting Pain Based on Physiological Data Using Random Forests and Score-Level Fusion

Dakshin Rathan

Abstract

This paper follows the development of a system of detecting pain using multiple random forests and score level fusion with physiological data. This has many important applications, including determining when a soldier is in pain when in combat. This paper will go over how the system works and the results that were obtained during the testing phase. The results show that while the model is a great first step, many improvements can be made to increase metrics like accuracy, recall, and precision. The paper will then go into some of the improvements that can be made to bolster the robustness of the model.

I. Introduction

In this paper, we are discussing a pain recognition system based on physiological data using score level fusion with random forest classifiers. This is an important development and accurately predicting whether someone is in pain is useful for soldiers in combat, children who cannot clearly communicate, and for patients in a hospital.

Below are some short discussions on related papers discussing pain recognition and different methods to achieve accurate pain recognition.

Paper 1: Improving Pain Recognition Through Better Utilization of Temporal Information

This paper explores the most accurate method of detecting pain through a video. Since videos have large file sizes, compression is done for temporal signal scanning. However, by using the spatial signals, a more accurate result is found.

Paper 2: Automatic Pain Recognition from Video and Biomedical Signals

This paper explores an automatic and continuous system of pain monitoring. It combines video analysis of facial expression and physiological data to make predictions using the BioVid Heat Pain Database.

Paper 3: Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities

This paper presents a pain recognition system using deep learning with multimodal data. It uses thermal data, video data pixel by pixel, and depth data for spatial analysis.

Paper 4: Automatic Recognition Methods Supporting Pain Assessment: A Survey

This paper shines a light on various pain recognition systems and evaluates them via a survey. This paper also discusses the challenge of validating pain recognition results, as it is subjective and difficult to know if someone is actually in pain or not unless they answer themselves.

Paper 5: Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images

This paper discusses a deep learning CNN model that takes spatial and temporal data from videos as well as facial resolution manipulation to improve the accuracy of pain detection. This paper presents a super-resolution algorithm to manipulate facial video frames with different resolutions.

II. Method

A decision tree is an algorithm that can classify input data based on a series of decisions. A basic structure of a decision tree is below. The tree starts at the root node and makes splits, or decisions,

where the data will move further down the tree. When a pure classification is achieved (ex. Pain or No Pain), a decision is made.

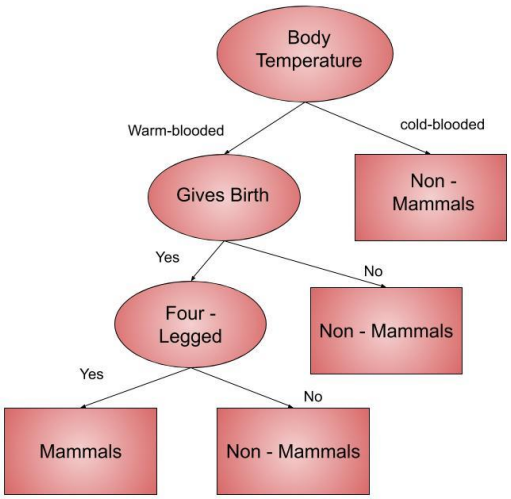

Figure 1 Decision Tree

However, a single decision tree is vulnerable to overfitting to the training data and bias of a single data set. To combat this in this paper, we use random forest.

A random forest randomly samples a portion of the input data with replacement and creates many different and unique decision trees. Because of this, different splits occur, and different decisions occur in each of the trees. After all the trees have made a decision, the random forest will make a classification in line with the majority decision among the trees.

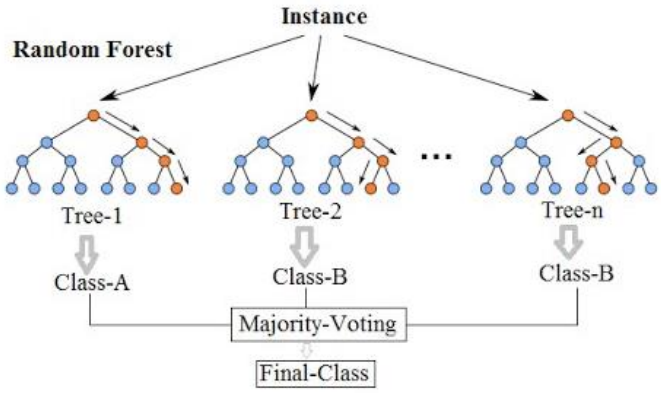

Figure 2 Random Forest

In this project, we classify whether a subject is in pain or not based on entries of signal data for four types of physiological signals: Respiration Rate (RES), Systolic Blood Pressure (SYS), Diastolic Blood Pressure (DIA), and electrodermal activity (EDA).

Training and Testing Random Forest

We used training and testing data of 30 subjects with 8 sets of physiological signal readings: 4 sets corresponding to each type of physiological signal when "pain" was reported and 4 sets when "no pain" was reported. To accomplish classification in this project, we trained four different random forests for each type of physiological signal. However, physiological signal entries are variable and random forests require each set of readings to contain the same number of features. To combat this, we down sampled the physiological readings of all entries to 5000 to create a uniform number of features. Then, we normalized each entry so values range between 0 and 1.

Because we are using one random forest for each type of physiological data, we filtered all the signal readings for each type of physiological data for each random forest. For example, all the RES training data will be filtered out to be trained on the RES random forest. This same process is done for the 3 other type of physiological data. The testing data is filtered out in the same way and is used as an input in the testing phase of the random forest. We generate a set of predictions for each subject in the testing data.

Score Level Fusion

After we get predictions from each random forest, we use a method called score level fusion to generate the final result for each subject. For each subject, we have a prediction from each of the 4 random forests corresponding to a type of physiological signal. Whichever choice (pain or no pain) has the majority vote among the 4 trees, is chosen as the final decision for a test subject.

III. EXPERIMENTAL DESIGN AND RESULTS

We used two different csv files as splits for training and testing. We also printed the confusion

matrix, accuracy, recall, and precision for each individual random forest and the combined prediction results after score level fusion for each alternative. The results of the experiment is below.

```
DiaPredictions
[[16 14]
 [15 15]]
accuracy:  0.5166666666666667
recall:  0.5
precision:  0.5172413793103449
SysPredictions
[[16 14]
 [14 16]]
accuracy:  0.5333333333333333
recall:  0.5333333333333333
precision:  0.5333333333333333
EdaPredictions
[[15 15]
 [14 16]]
accuracy:  0.5166666666666667
recall:  0.5333333333333333
precision:  0.5161290322580645
ResPredictions
[[16 14]
 [14 16]]
accuracy:  0.5333333333333333
recall:  0.5333333333333333
precision:  0.5333333333333333
```

Figure 3 Data1.csv: testing Data2.csv: training

```
accuracy:  0.5333333333333333
recall:  0.5333333333333333
precision:  0.5333333333333333
```

Figure 4 Results after fusion

```
[[14 16]
 [16 14]]
accuracy:  0.4666666666666667
recall:  0.4666666666666667
precision:  0.4666666666666667
SysPredictions
[[14 16]
 [17 13]]
accuracy:  0.45
recall:  0.43333333333333335
precision:  0.4482758620689655
EdaPredictions
[[14 16]
 [16 14]]
accuracy:  0.4666666666666667
recall:  0.4666666666666667
precision:  0.4666666666666667
ResPredictions
[[14 16]
 [16 14]]
accuracy:  0.4666666666666667
recall:  0.4666666666666667
precision:  0.4666666666666667
```

Figure 5: Data1.csv: training Data2.csv: testing

```
accuracy:  0.4666666666666667
recall:  0.4666666666666667
precision:  0.4666666666666667
```

Figure 6 Results after fusion

We can see that the accuracy, precision, and recall was very similar to when data1.csv was the testing data than vice versa. Another thing that we see is that the confusion matrices and statistics show that our model is not very accurate. Based on the data, we see that the accuracy, recall, and precision all hover around 50%, which is the normal guessing percentage.

There are several reasons why we think the model wasn't very accurate. The first reason is the size of the training data. Seeing that there were 5000 features fed into the random forest classifier, We don't think that 30 samples are enough data for the random forest to understand the importance or contribution of each of the 5000 features. In Project 1, where we had more favorable results (figure 7), we saw that even though there were only 30 samples, there were only 5 features for the random forest to work with, which might be more optimal.

A solution to this problem could be increasing the number of samples, which would provide more data for the model to make sense of the 5000 features. However, a drawback to this would be that it's difficult to get pain-related data from a lot of people. Another drawback would be that the training time would certainly take longer when there is more data to process and train on.

Another reason why we think the model wasn't accurate is due to the nature of the training and testing data. We see that data1 is data from strictly females and data2 is from males. Considering that there are plenty of physical differences between males and females, using training data for one gender to predict the pain status of another gender doesn't make a lot of sense. Many of the patterns and signals that the random forest picked up on during training won't necessarily have the same application when testing. If we were to use the female data for training, the model may have overfitted to the female pain response, which can become very unreliable when predicting the male pain response.

```
Conf. Matrix:
 [[2.4 0.6]
 [0.7 2.3]]
Accuracy Score: 0.7833333333333332
Recall Score: 0.7666666666666667
Precision Score: 0.8266666666666665
```

Figure 7 Project 1 results

IV. Discussion and Conclusion

We think physiological data is good for pain recognition. However, some physiological responses are shown to contribute to pain more so than others.

According to the data, the majority voting fusion approach did not improve the accuracy, recall, or precision by very much. Though we cannot conclude anything definitively, we can see that by the results below from our project 1 algorithm using the project 1 fusion method with hand crafted features, that the method results in higher accuracy, recall, and precision. However, it is important to note that there were several differences between this project and project 1 to the point that we don't believe that the higher accuracy is a result of using a different fusion method. We think that score level fusion (or majority voting fusion) was actually an improvement over the project 1 way of doing fusion. By separating different types of readings, the random forest can accurately understand patterns in a specific type of data. In other words, we think that if we want to predict pain from a diastolic blood pressure reading, then a random forest that is focused on only diastolic blood pressure is better than a random forest that generalizes itself to all types of readings. To improve on this method, we think that some knowledge of which types of readings contribute more to pain detection. If we have some ideas about this, then we think that weighting votes proportionally more for types of readings that more closely predict pain. Currently, there is majority voting among the results of all four random forests when choosing the final prediction. However, the method we are proposing is to give some categories more weight than others in deciding the majority decision. For example, if blood pressure is more closely related to pain than respiration is, then weighting blood pressure higher than respiration would make sense in our view.

We think that contributing other modalities can improve the accuracy of the machine learning models. In the papers we referenced in the introduction section, using data from thermal readings, facial photos, videos of different resolutions, and temporal and spatial analysis of those videos allowed for more reliable and accurate predictions. This allows the model to get a fuller picture of the different variables that contribute to accurately detecting pain. We think that adding these other modalities as separate types of readings to the score level fusion method would increase the accuracy of the model.

While this model was a forward step in pain detection using random forests with physiological data, we can improve much more. To reiterate, we have a few suggestions for how we can go about this. This includes increasing the number of samples used in training, separating the training and testing process by gender, using a weighted voting fusion method, and adding more modalities to the weighted voting fusion method.

References

L. Breiman, “Random forests,” *Machine learning*, vol. 45, 2001.

M. A. Haque, R. B. Bautista, F. Noroozi, K. Kulkarni, C. B. Laursen, R. Irani, M. Bellantonio, S. Escalera, G. Anbarjafari, K. Nasrollahi, O. K. Andersen, E. G. Spaich, and T. B. Moeslund, “Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities,” *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018.

M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T. B. Moeslund, P. Rasti, and G. Anbarjafari, “Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images,” *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, pp. 151–162, 2017.

P. Lucey, J. Howlett, and J. Cohn, “Improving pain recognition through better utilisation of temporal information,” *International conference on auditory-visual speech processing*, vol. 2008, 2008.

P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, “Automatic Pain Recognition from Video and Biomedical Signals,” *2014 22nd International Conference on Pattern Recognition*, 2014.

P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. Picard, “Automatic Recognition Methods Supporting Pain Assessment: A Survey,” *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.