# Noisy and Clean data

There are a number of factors that would affect the performance of the Decision Tree Algorithm, purity of dataset is definitely one of the major issues to be considered. Within this coursework, there are two sets of data provided, clean data and noisy data. The clean data have been expertly filtered, whereas the noisy data were obtained by individual feature GentleBoost templates built from Gabor wavelet features with particular filter scheme(Valstar and Pantic, 2012). With the use of two different sets of data, the average accuracy of two datasets are shown below:

|  |  |  |
| --- | --- | --- |
| Dataset | Clean | Noisy |
| Accuracy rate | 0.727 | 0.597 |

Table 1: Accuracy rates against Clean and Noisy Datasets

The accuracy rate of clean dataset is 72.7% while the noisy dataset is 59.7%. Although the accuracy rate is different from algorithms, in this decision tree algorithm, the reason that noisy data affects the accuracy is unique. While training noisy data, it may be the case where two contradictory expressions have similar AUs which makes the model itself has difficulty while identifying these contradictory expression(Zhu and Wu, 2004). In more details, table 2 gives more details of performance between different expressions:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Expression | Anger | Disgust | Fear | Happiness | Sadness | Surprise |
| F-measure in Clean Data | 0.57 | 0.737 | 0.724 | 0.852 | 0.598 | 0.821 |
| F-measure in Noisy Data | 0.29 | 0.688 | 0.579 | 0.736 | 0.484 | 0.71 |

Table 2: F-measure rates for six expressions with clean and noisy data models

From table 2, result of analysing expression Anger performs worst with only 57% accuracy in clean data set, the situation becomes even worse with rate of 29% in noisy data set. The similar case is Sadness, but with better performance in noisy dataset compared to Anger. The others have at least more than 50% of recognition rate no matter with clean or noisy dataset.

To deal with this issue, Brodley and Friedl has illustrated a general algorithm which creates classifiers with voting and consensus mechanisms in order to filter out dataset before training(E.Brodley and A.FRIEDL, 1999). This method is a fairly straightforward approach to deal with mislabeled training data; another approach has been implemented within the training process, cross validation. This coursework uses 10-fold cross validation to split datasets into sub-groups to enhance recognition rate.

[1] Valstar, M. and Pantic, M. (2012). Fully Automatic Recognition of the Temporal Phases of Facial Actions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(1), pp.28-43.

[2] Zhu, X. and Wu, X. (2004). Class Noise vs. Attribute Noise: A Quantitative Study. Artificial Intelligence Review, 22(3), pp.177-210.

[3]E.Brodley, C. and A.FRIEDL, M. (1999). Identifying Mislabeled Training Data. Journal of Artificial Intelligence Research, [online] pp.131-167. Available at: http://jair.org/media/606/live-606-1803-jair.pdf [Accessed 4 Feb. 2018].