CSE572 – DATA MINING

ACTIVITY RECOGNITION PROJECT: USING EMG DATA TO DISTINGUISH EATING ACTIVITY FROM NON-EATING ACTIVITY

Introduction:

Data mining utilizes knowledge from different domains (such as Database Systems, Machine Learning/Artificial Intelligence, Mathematics and Statistics) to identify patterns and useful information from 'big data'[1-3]. Data mining can involve preprocessing stage, execution of mining algorithms core methods and typically a third stage of postprocessing for visualizing the results of analysis[3]. A more comprehensive way to look at data mining, is to have the data mining process broken down into eight steps – defining the problem, data integration, data selection, data cleaning, data transformation, data mining, pattern evaluation and knowledge presentation[2].

The first step of defining the problem, entails your knowing why you are carrying out the data mining process - what is the aim/objective? According to Damien (2019): the main goal of the data integration step "is to offer the users a unified perspective of the data irrespective of whether they are derived from single or multiple sources"[2]. In this second step, you need to identify all possible sources of data, collect the data and merge the data. Because you now have a large amount of data that include a big chunk of irrelevant data, you want to select only those that are relevant. So in the data selection step, you want to select the data to form your target dataset. This target dataset is what you will subject to preprocessing[2].

The preprocessing stage consists of data cleaning and data transformation. In the data cleaning step, you want to clean the target dataset of errors, inconsistencies and inaccuracies before it can be subjected to other data mining techniques. In the data transformation step, you want to convert the target dataset into a "destination data", which is a format that is usable and recognizable to data mining tools/techniques. Some of the transformation techniques include smoothing, binning, clustering, regression, aggregation, normalization and generalization[2].

The data mining step involves using mining algorithms such as classification, regression, clustering, association rule mining and outlier detection to determine the pattern, relationship, and correlation within and among the datasets [2, 3]. This step is considered the core component of the whole data mining process. The choice of the algorithm used is mostly dependent on what the objective is for performing data mining [2].

The pattern evaluation and knowledge presentation steps form the post-processing stage, which involve knowledge filtering, interpretation and explanation, evaluation and knowledge integration[2, 4]. The pattern evaluation step involves evaluating how interesting the parameters/measures are, so that patterns that are truly interesting, impactful or relevant to give useful information are determined. Interpretation of the outcome of this step marks the transformation of data into a large bag of knowledge that can be used for making informed polices and business decisions[2]. The knowledge presentation step involves presenting (best through the use of visualization techniques) knowledge derived from the evaluation and interpretation of data mining patterns to the relevant stakeholders [2]. The post-processing stage can also involve checking the new knowledge for potential conflicts with previously induced knowledge. Several criteria used for the purpose of this stage include classification accuracy, understanding, computational complexity, and so on[4].

For my data mining project, I utilized data mining techniques to classify patterns in signals from EMG sensors into eating and non-eating activities. The problem definition, data integration and data selection steps were already accomplished before the project assignment. So what I obtained were target datasets. Based on the target datasets, I executed the preprocessing stage (data cleaning and data transformation/extraction steps), the data mining step (predictive mining algorithms - classification) and postprocessing stage (classification accuracy)

Explanation of the Solution

Target Data Set: My target datasets came from 30 users: 09, 10, 11, 12, 13, 14, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 36, 37, 38, 39, 40 and 41. For each user, there were two files containing "ground truth" information for fork and spoon activities respectively. These ground truth files contained information for identifying eating and non-eating activities in respective sensor data files. For each user, there was two types of sensor data (IMU and EMG); and for each sensor, there were two files – fork and spoon data files. For my project, I utilized the EMG data.

Preprocessing - Data Cleaning: The ground truth file consisted of three columns, the first column is the start frame of an eating action and the second column is the end frame. Each row is an eating action. To generate my "destination data" (the rows in the EMG files represented the eating and non-eating activities) from my target datasets, I multiplied the first and second column of the ground truth file by 100 and divided by 30. I then utilized the result of the first columns to identify the corresponding row in the EMG data that indicates the start of an eating action. Then I utilized the result of the second column to identify the corresponding row in the EMG data that indicates the end of an eating action. After extracting the eating rows, the rows that did not fall into these ranges (first and second column) were the non-eating rows. I then labelled the eating datasets by introducing a column containing "eating" and the non-eating datasets with column of "non-eating". At the end of these stage I had two (eating and non-eating) destination data for each user. Each destination data consisted of a label column and eight columns of EMG attributes. Since there were more rows in non-eating data, I took the first sets of eating rows that equals the number of rows in the eating data. So for each user I had equal number of rows for eating and non-eating datasets.

There were two users (18 and 25) that had more ground truth information than corresponding EMG data for fork and spoon respectively. These led to the generation of NAs by the data cleaning code. These NAs will obstruct downstream data mining steps if not removed. As explained earlier in the introduction, errors, inconsistencies and inaccuracies are characteristic of raw target datasets, and part of the data cleaning step is to get rid of these errors, inconsistencies and inaccuracies. So I removed all these erroneous rows of NAs from the users 18 and 25 datasets.

Preprocessing - Data Extraction/Transformation: There is an assumption that data is not randomly scattered but consists of subspaces that effectively represent the data. Feature extraction techniques can be used to transform or project a space of many dimensions into an alternative representation of fewer dimensions [5, 6]. This involves creating a new reduced set of features that summarize most of the information in the original set of features. Feature extraction has the advantage of reducing overfitting, improving accuracy, speeding up training, improving data visualization and enhancing the explanation of models [5]. For this step of the project, the five feature extraction methods I used were independent component analysis, mean, median, standard deviation and max. For each user, I concatenated my non-eating data to the eating data and then implement the five feature extraction methods above. This led to my generation of a reduced data set of five attributes instead of the initial eight EMG attributes.

I then proceeded to transform the data further by reduction of the feature space and keeping only those features which show the maximum distance between the two classes (eating and non-eating). For this purpose I used Principal Component Analysis technique.

Data Mining: The purpose of my project was to predict eating action based on EMG attributes. Since this is a classification problem, I used three predictive data mining (supervised) machine learning (ML) algorithms - decision trees (using rpart library in R), support vector machines (using e1071 library in R), and neural networks (using neuralnet library in R). For data mining, I carried out two type analysis. The first was user-dependent, in which I carried out each ML for each user. In this analysis I split my data into two and used 60% for training and 40% for testing. My training and testing data contained equal amounts of eating and non-eating rows. The second analysis was user-independent, in which I combined the first 18 user's data (i.e. 60% of users) for training and combined the remaining user datasets for testing. Note, that because of the limited time frame, I was allowed to use a subset of each user's data in the user-independent analysis. Even with the subset, the user-independent dataset was still quite larger (about 4:3) than the largest user-dependent datasets.

Postprocessing - Pattern Evaluation and Knowledge Presentation: I utilized three classifications accuracy metrics (precision, recall and F1 score) for pattern evaluation. For knowledge presentation, I utilized tables and plots.

Description of the Results:

For the user-dependent analysis, the average of accuracies (data not shown) ranged from 0.62 to 0.69. The neural networks produced the best average precision and F1 score, while decision trees produced the best average recall. For the user-independent analysis, the neural networks performed exceptionally well with the accuracies ranging from 0.94 to 0.98. On the other hand, the decision trees and support vector machines had accuracies that ranged from 0.50 to 0.57.

In conclusion, it is interesting to see that the average accuracies obtained using the user-dependent analysis was better for decision trees and support vector machines than with the user-independent analysis. However, the opposite was the case with neural networks. This very much agrees with the concept that more complex algorithms (e.g. neural networks) perform better with larger datasets (low bias and high variance), while relatively simpler once perform better with smaller datasets. It is also noteworthy that with the user-dependent analysis, certain users had better metrics than others and this was consistent across the three algorithms. These differences may be due to the quality of the data, as low quality and mislabelled data may result in defective machine-learning systems [7]. The source of the low quality could be because:

1. There could be some errors, inconsistencies and inaccuracies in the given target datasets. The lack of enough EMG data in two user data (18 and 25) compared to the ground truth is evidence of such. With these two datasets, the errors were discovered and corrected. However, other errors may be hidden.
2. Errors may also have been introduced during data capturing.
3. The accuracies did not correlate with data size (data not shown), so the accuracies were not size-dependent.

Description of My Contributions to the Project:

The project was not a group project, so I executed all the steps of the project by myself by following the instructions given by the course professor and the teaching assistants. I also did all the write up by myself.

New Skills, Techniques, or Knowledge I Acquired from the Project:

The new expertise I obtained through this project include ability to:

a. Understanding and executing the steps necessary to solve a data mining problem.
b. Carry out independent component analysis using fastICA library in R
c. Executing two of the three machine learning algorithms - decision trees (using rpart library in R) and support vector machines (using e1071 library in R).
d. Although I have previously executed neural networks, in this project, I learnt how to use the popular "rules of thumb" for making decisions on the number of neurons for the hidden layer(s) [8].

References

[1] J. F. Dipnall, J. A. Pasco, M. Berk, L. J. Williams, S. Dodd, F. N. Jacka, and D. Meyer, "Fusing Data Mining, Machine Learning and Traditional Statistics to Detect Biomarkers Associated with Depression," *PLoS ONE,* vol. 11, no. 2, pp. e0148195, 2016.

[2] L. Damien, "DATA MINING: Your Ultimate Guide to a Comprehensive Understanding of Data Mining," Kindle Edition, 2019.

[3] C. Bellinger, M. S. Mohomed Jabbar, O. Zaïane, and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC public health,* vol. 17, no. 1, pp. 907-907, 2017.

[4] J. Diaz-Arevalo, M. Herrera, J. Izquierdo, and R. Pérez-García, *The tasks of pre and post-processing in Data Mining applied to a real world problem*, 2010.

[5] P. P. Ippolito, "Feature Extraction Techniques," 2019.

[6] G. D. Clifford, "BLIND SOURCE SEPARATION: Principal & Independent Component Analysis ", 2008.

[7] S. Udeshi, X. Jiang, and S. Chattopadhyay, "Callisto: Entropy based test generation and data quality assessment for Machine Learning Systems," 2019].

[8] J. Heaton, *Introduction to Neural Networks for Java, 2nd Edition (2nd. ed.)*: Heaton Research, Inc. , 2008.