E2I5 année 2021-2022

TP2

Traitement de données – Classification

Les données sont extraites d'un article et en partie modifiées pour le TP.

Luis M. Candanedo, Véronique Feldheim. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models, Energy and Buildings, 112, pp 28-39, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2015.11.071.

Voici le descriptif des données et de leur recueil (extrait de l'article) :

This research has used data recorded from light, temperature, humidity and CO2 sensors as a means to detect occupancy and a digital camera to establish ground occupancy for supervised classification model training. Combinations of these sensors can already be found in many buildings.

Data collection and setup

An office room with approximate dimensions of 5.85m × 3.50m × 3.53m (W × D × H) was monitored for the following variables: temperature, humidity, light and CO2 levels.
An additional feature/variable is the humidity ratio. The humidity ratio in kgw/kgda was calculated using the measured temperature and relative humidity.
A microcontroller was employed to acquire the data. A ZigBee radio was connected to it and was used to transmit the information to a recording station. A digital camera was used to determine if the room was occupied or not. The camera time stamped pictures every minute and these were studied manually to label the data. See Fig. 1 for photograph of the setup. The time stamp of the data has been exploited in this work by extracting the number of seconds from midnight for each day.

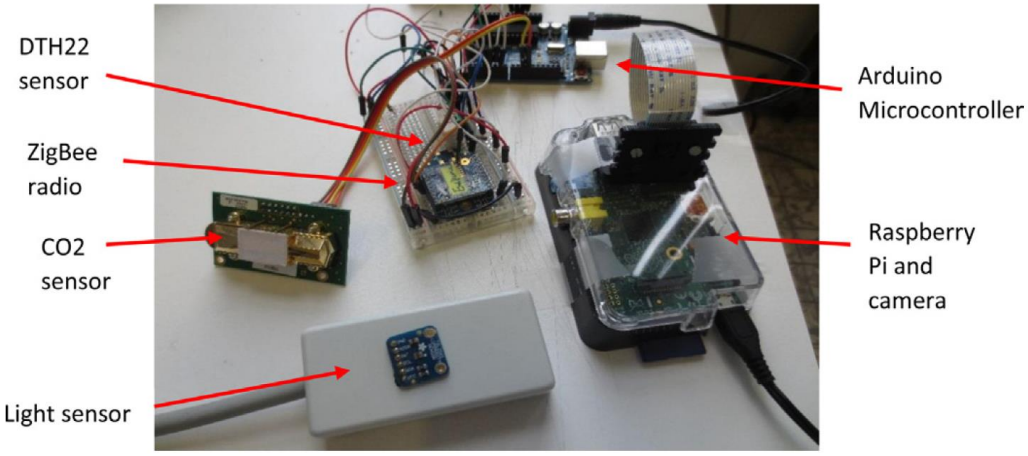

Fig. 1. Data acquisition setup showing the light, CO_2, DHT22 (temperature/humidity) sensors, a ZigBee radio and a microcontroller card and the digital camera controlled by a Raspberry Pi.

Environmental conditions

The data was recorded during winter in Mons, Belgium during the month of February. The room was heated by hot water radiators that kept the room above 19◦C. In order to estimate the difference in occupancy detection accuracy given by the models, they are tested for data sets when the office door is open and closed. The readings were recorded at time intervals of 14 s or 3 to 4 times per minute, and then averaged for the corresponding minute. The sensors were placed on a desk as shown in Fig. 2. The distance to the closest occupant was 1.1 m and to the second occupant about 2.9 m.

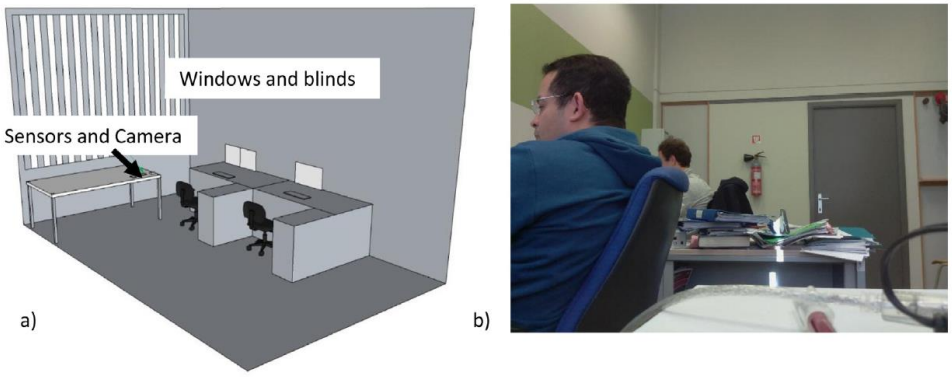

Fig. 2. (a) Room sketch showing the position of the sensors and the position of the occupants (b) Example of one of the pictures from the digital camera used to establish ground occupancy.

Les données sont enregistrées, extraites et mises en forme dans les différents fichiers excel.

Le but du projet est d'analyser les variables et de quantifier la possibilité de détecter la présence ou non de personnes dans une pièce.

Vous devez rendre un rapport détaillé et précis. Le rapport ne fera apparaitre en détails que les analyses pertinentes ; vous listerez également toutes les analyses que vous avez faites (pour que je puisse évaluer la quantité de travail réalisée et la réflexion que vous avez menée).

Vous devez finalement formuler une conclusion claire sur la nécessité d'un ou de plusieurs capteurs pour pouvoir détecter la présence de personnes dans un bâtiment (quels capteurs, justifier leur choix, dans quel type de bâtiment, le domicile d'un particulier, des bureaux etc.)

Voici une liste de questions non exhaustives pour vous aider dans votre démarche

1/ Vous utiliserez la base de données « training » pour toutes les 1ères analyses ; les deux bases de test seront utilisées pour évaluer les performances des classifiers et les comparer. Il est dans un premier temps important de décrire les 3 jeux de données : nombre d'observations, nombre de variables, nombre d'observations par classe, moyenne des variables par classes, boxplot des variables etc.

2/ Visualiser les données (training) sous la forme de courbes temporelles par exemple sur une journée (figure 3 de l'article).

3/ Effectuer une analyse discriminante (approche descriptive) de la base « training ».

4/ Effectuer également une analyse des différentes variables par des courbes ROC (toujours sur la base « training ».

5/ Choisir la « meilleur » variable et fixer un seuil sur cette variable pour la classification. Réaliser alors une classification d'une des bases de test en utilisant cette variable avec le seuil choisi ; faites de même pour la variable discriminante obtenue sur la base « training ».

6/ Vous allez maintenant réaliser une analyse discriminante linéaire selon une approche prédictive : apprentissage sur la base « training » et test sur les bases « test ».

7/ Proposer un classifier des k plus proches voisins (Attention vous devrez séparer la base d'apprentissage en une base d'apprentissage et une base de validation qui servira à trouver le k « optimal ».

8/ Comparer sur les bases de test : l'analyse discriminante linéaire, quadratique et les k-ppv.

9/ Discuter des résultats ; des différences entre les 2 bases ; analyser les échantillons mal classés etc.