DATA MINING FOR LABELLING DATA FROM SMART BUILDINGS

Kristina Redenšek [1], Cyril Oswald[2]

[1] University of Ljubljana, krisitnaredensek@gmail.com
[2] Czech technical university in Prague, cyril.oswald@fs.cvut.cz

Abstract: In recent years, the development trend of buildings is in to maximize user's comfort and to minimize energy consumption. As a result of these trends, smart buildings are becoming more and more popular, equipped with many sensors that measure parameters that are connected with the residents' comfort and the energy efficiency of the building, today there is a flood of wireless sensors that can be easily installed even into an already existing building. As a result of a lot of buildings equipped with sensors it often happens that we get some data from sensors that is without labels. However, knowing the labels in further processing is crucial. Manually determining labels is time consuming work. Thus, our goal was to create a program which can categorized a set of data to one of multiple predefined category. A method that suited our task the most was a method with a decision trees. In order to use this method we did first preprocessing of data to choose the most appropriate features and after classification. Result of our work was a program that included preprocessing data and data classification using the decision trees method. Predictions were in more than 90% correct for all classes. To conlclude, this program greatly facilitates and accelerates data labelling, which was for before done manually.

1 Introduction

In recent years, the development trend of buildings is in to maximize user's comfort and to minimize energy consumption. As a result of these trends, smart buildings are becoming more and more popular, equipped with many sensors that measure parameters that are connected with the residents' comfort and the energy efficiency of the building. If over the past years, the focus has been on installing sensors in buildings, today, there is a flood of wireless sensors that can be easily installed even into an already existing building. The result of a lot of sensors is a large amount of data. If we want to benefit from measuring all possible quantities, we need to use this data wisely. Therefore, in recent years, a great emphasis has been placed on the analytical treatment of these data. If data is properly captured and statistically processed, we can make excellent statistics models that dynamically regulate the operation of HVAC and other systems and thus minimize energy consumption and increase the comfort in buildings. The most popular approach in field of analysis of data is data mining. One of the most common methods that are used are: classification, clustering, regression, association rule learning. With these methods we can find some typical patterns and relations in the data.

As a result of a lot of buildings equipped with sensors it often happens that we get some data from sensors that is without appropriate labels. However, knowing the labels in further processing is crucial. Manually determining which data do we have is at the huge amount of data time demanding work. Thus, our goal was to create a program which can categorized a set of data to one of multiple predefined category. Doing that, we wanted to ensure a high percentage of accuracy in the implementation of this process. As the most suitable method we saw classification method of data mining. For getting better results we paid special attention in phase of preprocessing data, in which we also used visualization of data to have a better image of data properties.

The rest of paper is organized as follows. Section 2 some elements of data mining are introduced. Section 3 describes methodology of work. Section 4 reveals result of our work discuss it. In section 5 finally conclusion is made.

2 Data mining

Data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Some of the methods involved in computing process of data mining are at the intersection of machine learning, statistics, and database systems. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns. Examples include linear equations, rules, clusters, graphs, tree structures, and recurrent patterns in time series. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Typically deals with data that have already been collected for some purpose other than the data mining analysis. This means that the objectives of the data mining exercise play no role in the data collection strategy. This is one way in which data mining differs from much of statistics, in which data are often collected by using efficient strategies to answer specific questions. For this reason, data mining is often referred to as "secondary" data analysis [1, 2]. Most common classes of tasks supporting these steps are: anomaly detection (outlier/change/deviation detection), association rule learning (dependency modelling), classification, regression, summarization [3].

2.1 Decision trees

A “divide-and-conquer” approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree. Nodes in a decision tree involve testing a particular attribute. Usually, the test at a node compares an attribute value with a constant. However, some trees compare two attributes with each other, or use some function of one or more attributes. Leaf nodes give a classification that applies to all instances that reach the leaf, or a set of classifications, or a probability distribution over all possible classifications. To classify an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached the instance is classified according to the class assigned to the leaf [4].

Metrics

Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items. Different algorithms use different metrics for measuring "best". These generally measure the homogeneity of the target variable within the subsets. These metrics are applied to each candidate subset, and the resulting values are combined to provide a measure of the quality of the split [5].

Gini impurity

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. Gini impurity can be computed by summing the probability p_i of an item with label i being chosen times the probability $\sum_{k \neq i} p_k = 1 - p_i$ of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category [6].

To compute Gini impurity for a set of items with J classes, suppose $i \in \{1,2,\ldots,J\}$, and let pi be the fraction of items labeled with class i in the set.

$$I_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1 - p_i) = \sum_{i=1}^{J}(p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2 = 1 - \sum_{i=1}^{J} p_i^2 \quad (1)$$

3 Methodology of work

Our goal was to create a program which can classify un unknown set of data to one of multiple classes. A method that suited our task the most was a method with a decision tree. For making a classification program, we received a piece of data from a data warehouse. The data was collected in an office building somewhere in Czech Republic in January 2017. The variables were: indoor air temperature, outdoor air temperature, supply water temperature, return water temperature, warm service water temperature, warm service circulation, pump operation, energy meter actual and energy meter cumulative. This data was already manually classified so we could use it as a training data.

3.1 Work flow

We can divide the process of our program into two steps. First step was preprocessing of data. This is a step that we need to do before starting an analysis of data and it is a very important step in doing analysis, because the quality of our later work (data mining) with this data depends on a quality of input data. That is the reason for paying a lot of attention at this step. Second step was data mining of data.

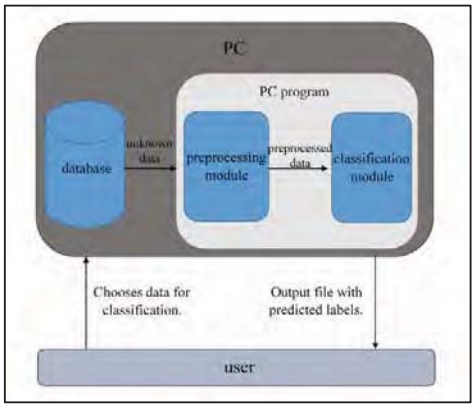

Fig. 1.: Block schematic presentation of a program.

3.2 Preprocessing

In preprocessing step, a target data set was assembled. As data mining can only uncover patterns actually present in the data, the target data set must to be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. To find out important feature of data sets we first visualized data in plots. After that we decided to consider next features: variable type (*boolean* or non *boolean*), measures of descriptive analysis (mean, maximum, minimum, standard deviation) and order of data (ascending, descending). We decided to drop missing data.

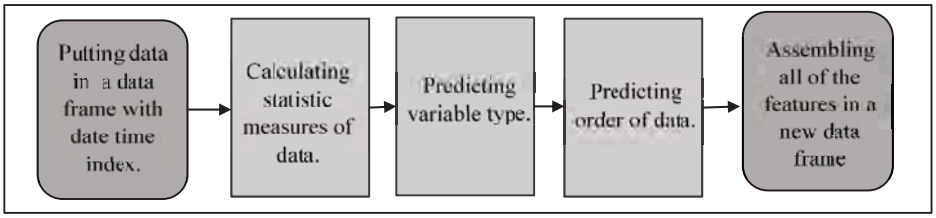

Fig. 2.: Flowchart of a preeprocessing stage.

3.3 Classification

Tab. 1.: Classes and matching labels.

Class	Labels
indoor air temperature	indoor air temperature
outdoor air temperature	outdoor air temperature
water temperature	supply water temperature
	return water temperature
	warm service water temperature
energy meter actual	energy meter actual
energy meter cumulative	energy meter cumulative
boolean data	pump operation
	warm service water circulation

After making a visualization of data we noticed that some of data had very similar shape and accordingly features, so it was hard to distinguish between them. Due to that, we decided that we join some label into one class. In a table 1 determined classes and labels that match a class can be seen.

We did tests and according to them we made first separation of classification manually in order to get better results. So we implemented an algorithm that separated *boolean* and non boolean data. Since the *boolean* data was

already a class for classification, we did not process this data anymore. For the other data further mining was needed.

We did a classification with a decision tree with this data. For executing this method, we used library scikit – learn. First we assembled training data to make a decision tree. The measure of the quality of the split set in a decision tree was Gini impurity. For keeping the decision tree understandable and not to complicated, we limited the depth of a decision tree to 5.

4 Results

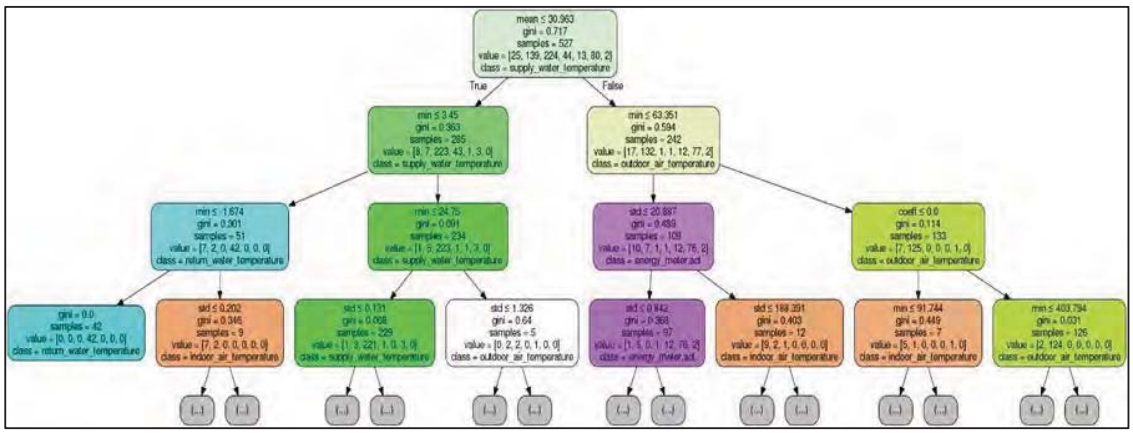

Fig. 3: A Classification tree for classifying non boolean data.

In Figure 3 is a decision tree that was made for classifying non *boolean* data into classes. As expected after seeing plots the first separation criteria were mean value of variables. In second step min values were obtained and in further steps the other features. Results of classification in percentage can be seen in Figure 4.

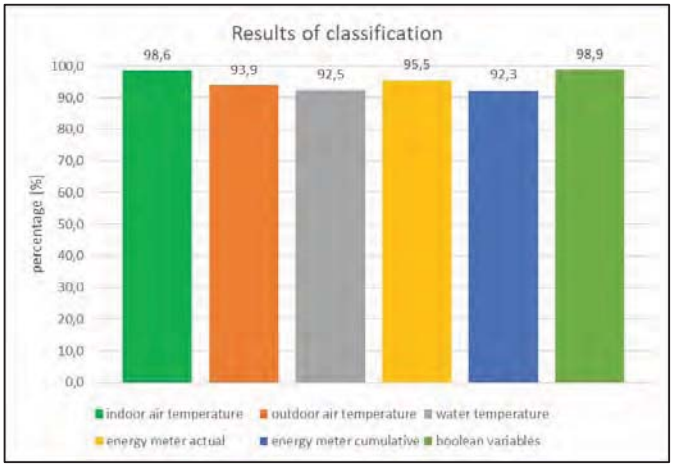

Fig. 4.: Results of classification in percentage.

Percentages of correct prediction of class are encouraging. All prediction results were in more than 90% correct. The best results were for *boolean* variables (98,9 %) and for indoor air temperature (98,6%). The worst results were for energy meter cumulative (92,3%), which is a surprise according to its unique shape of plot. A reason for that result are probably outliers such as always 0 value or lower values than usual. The result for water temperature is also one of the worst ones, even though we made a common class for 3 variables.

5 Conclusion

Our goal was to make a program that will be able to label a set of data. In order to fulfill this task, we made:

- A program that includes preprocessing data and data classification using the decision tree method;
- Preprocessing part of the program which takes care of data cleaning and computation of parameters of data for classification;

- Classification part of the program, which uses training data to build a decision tree and according to this decision tree than classifies tested data;
- A program that greatly facilitates and accelerates data labelling, which is for now done manually.

For classification, only those classes, which gave us satisfying results that are still reasonable and suitable for further use, were used. Thus, some classes included several variables. With this approach we achieved in more than 90% correct prediction for all the set classes.

In future program can be improved. One of the biggest improvement would be an algorithm that can classify all the variables with high enough accuracy. To accomplish that a deeper look at data is needed to find some specific features for each variable. Also other methods as regression and mutual information could be used.

References

[1] Hand, D; Mannila, H; Smyth, P: *Principles of Data Mining.* A Bradford Book The MIT Press Cambridge, p. 5-20, 2001.
[2] KDD: *Data Mining Curriculum.* Available on: http://www.kdd.org/curriculum/index.html on 20. 12. 2017.
[3] Fayyad, U; Piatetsky-Shapiro, G;Smyth, P: *From Data Mining to Knowledge Discovery in Databases*. AI Magazine, 1996.
[5] Witten, I.H; Frank, E: *Data Mining (Practical Machine Learning Tools and Techniques). Second Edition.* Morgan Kaufmann, p. 62 – 82, 2005.
[6] Rokach, L; Maimon, O: *Top-down induction of decision trees classifiers-a survey.* IEEE Transactions on Systems, Man, and Cybernetics, Part C. 35 (4): 476–487, 2005.

Selected article from
Tento dokument byl publikován ve sborníku

Nové metody a postupy v oblasti přístrojové techniky, automatického řízení a informatiky 2018
New Methods and Practices in the Instrumentation, Automatic Control and Informatics 2018
28. 5. – 30. 5. 2018, Příbram - Podlesí

ISBN 978-80-01-06477-1

Web page of the original document:
http://control.fs.cvut.cz/nmp
http://iat.fs.cvut.cz/nmp/2018.pdf

Obsah čísla/individual articles:
http://iat.fs.cvut.cz/nmp/2018/