Building Data Classification and Association

Christine Sjölander & Anna Åberg

Almost half of the energy consumption i EU originates from heating and cooling buildings. With smart control and analysis systems the consumption can be cut by thirty percent. In this article, attempts to automatically connect buildings with such smart systems are described. The results are promising, but much remains to be done.

A commercial building may contain thousands of signals, such as temperature and air-flow measurements, which are managed and controlled by a Building Management System, BMS. The BMS also keeps track of metadata such as names of the signals within the system. In order to minimise energy consumption Building Analysis Systems, BASs are used for buildings globally. For the BAS to be deployed it must connect to the BMS. This connection process is today performed manually which makes it both time consuming and error prone.

In order to connect the BAS and BMS the *type* of the signal and what subsystem, or *equipment* it belongs to must be known. The problem is therefore split into the *classification problem*, finding the type, and the *association problem*, finding the equipment. In Figure 1 the optimal solution is depicted. Time series data from a signal in the BMS is inputted to the solution which outputs the type and the equipment of the signal.

The solution was implemented using machine learning methods. The basic idea behind machine learning is that a programmer designs a framework for the algorithm and provides example input data with desired answers. The computer then finds connections between the input and answers by minimising the difference between the output from the model and the answers. More specifically, feature based machine learning were used. It means that instead of inputting raw time series data, the signals were described using different measurements called *features* which were then inputted to the algorithms.

For the classification problem statistical features such as mean and maximum value, the dominant frequencies and information about the gradient were extracted. For the association problem, on the other hand, the aim was to find features that could measure how connected two signals were. Therefore information on when events occurred and the dominant frequencies was extracted from the measured data. Signals were later compared by taking the difference between these measurements and this was used as input to algorithms. The measured similarity between different signals' names and paths, i.e. strings containing information of where in a system signals were located, were also used.

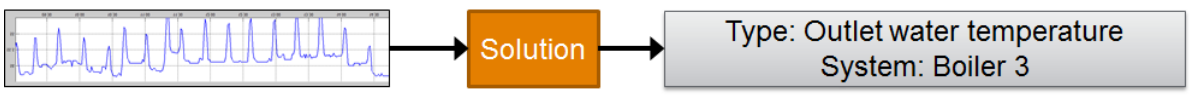

Figure 1: An overview of the idea behind the problem solution.

Table 1: Overview of the overall best performing models for each method on the classification problem.

Best results for the classification problem				
Data set	**Method**	**Accuracy top one /%**	**Accuracy top five /%**	**Accuracy diff /percentage points**
A	Random forest	100	-	0.0
	Gradient boosting	100	-	0.0
	Neural network	99	-	0.4
B1+2	Random forest	85.0	98.8	9.5
	Gradient boosting	85.0	98.1	10.3
	Neural network	75.0	94.5	-11.6
B2	Random forest	63.2	88.9	29.6
	Gradient boosting	60.2	90.7	33.2
	Neural network	72.8	92.0	-1.5
C	Random forest	59.8	92.9	10.3
	Gradient boosting	61.0	93.1	7.0
	Neural network	57.0	92.7	2.8

For this project three data sets were available. *Data set A* was the smallest and also the least complex data set. It consisted of five signal types, each measured in 51 rooms. *Data set B* was slightly more complex with sixteen different signal types measured in two buildings. This allowed for evaluation of how well solutions performed on data from a building that they had not trained on. The signals in Data set B could not be used for the association problem as equipment labels were missing. The last data set, *Data set C*, was the most complex data set. Signals in this set were distributed over 51 different classes and 5 equipment types.

In Table 1 the best results for the different machine learning methods implemented for the classification problem are presented. For Data set A, the simplest data set, the models achieve over 95% accuracy, i.e. in over 95% of the tested examples the algorithms' most probable guess was correct. For Data set B, the one with two buildings, there are two results presented as attempts were made for training and testing on both buildings, B1+2, and training and testing on different buildings, B2. For B1+2 the accuracy were 75-85% with gradient boosting as top performer, but for B2 the accuracy dropped to 63-73% with neural network as the best method. Clear was also that the accuracy difference for random forest and gradient boosting models were high. An accuracy difference of e.g. 10 percentage point means that the model has a 10 percentage points higher accuracy for data it has trained on compared to data it was tested on. This is called overfitting and means that the models have adjusted to patterns specific to the training data, such as noise. The overfitting was most severe for Data set B2 where it reached 33% for gradient boosting. However, some neural networks did not differ very much in performance between Data set B1+2 and B2. Neither did the best of them overfit to any great extent. Finally, the most complex data set, Data set C, was explored. Top one accuracy was only about 60% with gradient boosting models being top performers. The accuracy for the top five guesses, however, remaind high with a score of close to 93% for all methods.

Unfortunately 80% accuracy for the top guess was not reached for all data sets as required for a fully automated solution. It was however reach for the top five guess accuracy meaning that these models were at least good enough to be used for a semi-automated solution. Though the overiftting seen for several data sets can pose a future problem and should be taken seriously, the accuracy scores suggest machine learning on time series data to be a viable path. Further on, the results demonstrate issues as models were presented with data from a building they had not trained on. This can likely be helped by training on several other buildings to create a more generalised behaviour.

For the association problem the results can be found in Table 2. For Data set A, support vector machines, SVMs, were implemented to perform pairwise comparison. Since the same five signals

Table 2: Overview of the overall best performing models for each method on the association problem. For SVM the average results for all best model is presented.

Best results for the association problem			
Method	**Data set**	**Accuracy /%**	**Wrong per right**
SVM with path and name	*A*	98.0	0.4
SVM without path and name	*A*	77.0	4.5
String comparison	*C*	56.0	1.6

were measured in each room of the set it was possible to design four SVMs where each SVM was trained on deciding if signals of two specific types belonged to the same equipment. Meaning that the first SVM was trained on comparing signals of type one and two, the next SVM on type one and three and so fourth. The results presented in the table are the averages of these models. As seen in Table 2, the accuracy was close to 100% when signal names and paths were included while it dropped to 77% when this information was excluded. The ratio of incorrect guesses to correct guesses were also measured. Excluding the information in the signal names and paths also affects the ratio negatively as it increases from 0.4 to 4.5.

For the most complex data set, Data set C, a pure signal name comparison was used since the data set did not have the nice structure of Data set A. With more insight to the systems it is possible that a more intelligent design could be implemented for this data set as well. The string comparison method was not a machine learning algorithm, but simply measured the similarity of the strings and decided if the signals belonged together based on a threshold value. By tuning the threshold value a model with 56% accuracy and 1.6 wrong matches per right ones was found. No general conclusions could be reached from this as the implementation could only be tested on one building and naming conventions can differ between buildings.

The solutions presented directly above are not as feasible as those for the classification problem. The accuracy on Data set A was promising, but the solution with pairwise comparisons rely on a nice structure within the class/equipment-system which is not guaranteed for all BMSs. Such a solution will likely prove hard to implement on more complex systems. However, for both the SVMs and the string comparison it possible to conclude that the names and paths of the signals hold important information and should be taken into consideration in future work.