Classification algorithms in the material science and engineering data mining techniques

A Dardzinska\textsuperscript{1,2} and M Zdrodowska\textsuperscript{1}

\textsuperscript{1}Faculty of Mechanical Engineering, Department of Biocybernetics and Biomedical Engineering, Bialystok Technical University, Bialystok, Poland

Email: a.dardzinska@pb.edu.pl

Abstract. Data mining is an analytical process, which deals with the study of large data sets in search of patterns, correlations between data, and later their evaluation. The goal of data mining is usually prediction, among others sales volume, customer activities, extension ratios or the scale of customer loss. Data mining techniques allow finding previously unknown dependencies and schemas that can be used to support decision making or database description. Data mining techniques are developing very quickly and are more and more often used not only in typical fields such as customer relationship or management, but also in medicine, biomechanics, industry, materials sciences or mechanical engineering. The aim of this work is to evaluate the effectiveness of selected data mining techniques for predicting the concrete compressive strength, and to identify the features having the greatest impact on its compressive strength. The study analyzed the data of 1030 concrete samples using five known classification algorithms (C4.5, Random Forest, Naive Bayes Classifier, Supporting Vector Machine SVM) and neural networks (Multilayer Perceptron), which allowed to build an exploration model given with an accuracy of over 99\%. Potential features of concrete that may affect its compressive strength are also pointed out.

1. Introduction

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data, which uses many experiments and methods of artificial intelligence and machine learning. KDD is a complex process, which implementation involves the preparation of data, their exploration and interpretation of results [1, 2]. Most often the KDD process consists of the following steps: data selection, data transformation, exploration, that is extraction of knowledge from data, and interpretation of results. The major phase of KDD process is data mining, which use the proper algorithm for finding dependencies and schemas in the prepared data set, and then their representation in the understandable form [3-5]. The most popular forms of representation of the KDD are decision trees and logical rules [6-8]. Data mining uses many different techniques that build specific types of knowledge. Depending on the purpose of the discovered knowledge, it can map classifications, regressions, clustering, characteristics, discrimination, associations, etc. The most popular data mining technique is classification. Classification maps the data into the predefined classes and groups. It is used to predict group membership for data instances. There are many areas that adapt these techniques. For example, in medical database, we extracted the knowledge in form of rules, which help to classify individual diseases, and give the correct diagnoses [9, 10].

Due to the development of data mining techniques, it will achieve higher and higher popularity - not only in the typical areas (customer relationship, management), but also in other fields, such as medicine,
biomechanics, industry, materials sciences or mechanical engineering [11, 12]. Information systems have enormous data resources. However, this information becomes meaningful only when the correlations with other information can be determined. This is what data mining algorithms deal with, which allow finding new, significant and correlated information in the data set. They are used for classification of objects classify as well as for predictive purposes [12].

Concrete is the most important material in civil engineering. High-performance concrete is a highly complex material, which makes modeling its behavior a very difficult task. The concrete compressive strength is a highly nonlinear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate [13].

The aim of this study is to evaluate the effectiveness of selected data mining techniques for predicting the concrete compressive strength, as well as to identify those features of the concrete that have the greatest impact on its compressive strength. It will also be analyzed how the reduction of attributes describing samples affects the accuracy of the classification.

### 2. Methodology

For research we use concrete compressive strength data set from UCI Machine Learning Depository. Our dataset contains data of 1030 concrete samples from the investigations. Each sample is represented in the data set by 9 attributes – 8 quantitative input variables and 1 quantitative output variable, which are shown in table 1.

**Table 1. Attribute information [13].**

| Name                                | Data            | Measurement            | Description        |
|-------------------------------------|-----------------|------------------------|--------------------|
| Cement (component 1)               | quantitative    | kg in a m³ mixture     | Input variable     |
| Blast Furnace Slag (component 2)   | quantitative    | kg in a m³ mixture     | Input variable     |
| Fly Ash (component 3)              | quantitative    | kg in a m³ mixture     | Input variable     |
| Water (component 4)                | quantitative    | kg in a m³ mixture     | Input variable     |
| Superplasticizer (component 5)     | quantitative    | kg in a m³ mixture     | Input variable     |
| Coarse Aggregate (component 6)     | quantitative    | kg in a m³ mixture     | Input variable     |
| Fine Aggregate (component 7)       | quantitative    | kg in a m³ mixture     | Input variable     |
| Age                                 | quantitative    | Day (1–365)            | Input variable     |
| Concrete compressive strength      | quantitative    | MPa                    | Output variable    |

**Table 2. General parameters of data sets [13].**

| Name                                | Minimum | Maximum | Average |
|-------------------------------------|---------|---------|---------|
| Cement (component 1)               | 102.00  | 540.00  | 281.17  |
| Blast Furnace Slag (component 2)   | 0.00    | 359.40  | 73.90   |
| Fly Ash (component 3)              | 0.00    | 200.10  | 54.19   |
| Water (component 4)                | 121.75  | 247.00  | 181.57  |
| Superplasticizer (component 5)     | 0.00    | 32.20   | 6.20    |
| Coarse Aggregate (component 6)     | 801.00  | 1145.00 | 972.92  |
| Fine Aggregate (component 7)       | 594.00  | 992.60  | 773.58  |
| Age                                 | 1.00    | 365.00  | 45.66   |
| Concrete compressive strength      | 2.33    | 82.60   | 35.82   |

Table 2 presents the general parameters of concrete evaluated in this study [13].
All samples were classified into three groups: high (more than 56 MPa) concrete compressive strength (128 samples), medium (between 29 and 56 MPa) concrete compressive strength (527 samples) and low (below 29 MPa) concrete compressive strength (375 samples). In addition, the data was divided into two separate data sets: a set of training data (learning) to build the model (80%) and a set of test data for model evaluation (20%).

In this work classification was provided for three models:

- model 1 – all available variables (9 attributes)
- model 2 – variables extracted by means of attributes selection – chi-square test (5 attributes: age, cement, water, superplasticizer and coarse-aggregate)
- model 3 – new variables – the principal component analysis (PCA) - 3 new attributes.

All data mining algorithms have been applied using Weka Software (version 3.8.1., Machine Learning Group, University of Waikato, New Zealand), which as a tool in machine learning and knowledge acquisition allows for initial data processing, grouping, classification, regression, visualization, or discovery of association rules [3, 14]. Due to the high accuracy, five algorithms were used for the classification [1, 6 -8, 12]:

- J48 (C4.5) – the J48 tree is generated using the C4.52 algorithm, which divides the original data set by each of the variables. In this way, there are as many variants of the division as there are explanatory variables in the set. For each division, the value of the information gain metric is calculated, which is defined as the entropy increase in each of the subsets. The variable with the highest information gain becomes the first node of the tree. Then, this operation is repeated for all members until all instances are exhausted. The course of the process can be represented in the following steps: the data set E should be divided by each variable and the value of information gain, i.e. the increase in entropy of the obtained subsets compared to the original set, should be calculated; one should choose the variable a, which ensures the highest increase in information, and divide the original set according to it; the operation should be repeated on each subset until the instance is exhausted.

- Random Forest – two and multi-class classification algorithm. The operation of random forests is based on classification using a group of decision trees. The algorithm begins its operation by building many decision trees (the number of trees is defined by the user). For each tree a random observation attempt is selected, consisting of several explanatory variables (the number of variables in each tree is the second parameter defined by the user). Based on the maximization of information profit mechanism, further attributes are selected for distribution. The final decision is made as a result of majority voting on the classes indicated by individual trees.

- Naive Bayes Classifier – two-class and multi-class classification method. It was based on Bayes' theory of category prediction in an unknown dataset. It assumes complete independence of individual variables (which is rarely reflected in the real world). Hence its name - a naive classifier. Its operation is based on the simple conditional probability of events. We include it in one of the simplest algorithms. Despite its simplicity, it can give better results than more complex classification methods.

- Support Vector Machine (SVM) – the SVM method is a relatively young method for solving regression and classification problems. Classification using the SVM method is carried out similarly to artificial neural networks using a training set. It is an excellent alternative to artificial neural networks and often the correct classification is better for SVM. The essence of the SVM method is the construction of an "optimal hyperplane", whose task is to separate data belonging to opposite classes with the greatest possible margin of confidence.

- Multilayer Perceptron (MLP) – Two and multi-class classification algorithm. Perhaps the most sophisticated algorithm of all, inspired by the action of the human brain. It also applies to regression problems. The operation pattern of the neural network is described using an acyclic directed graph. The main element of the neural network is the processing neuron. There are many neurons in the network that have any number of inputs and outputs. Neurons are grouped into layers in which each neuron is connected to each neuron of the preceding layer.
values are therefore transferred progressively between individual layers of the neural network. In the next layers, operations are performed on variables until the result value is reached at the end of the graph.

The ACC (Total Accuracy) measure was used to assess the above classifiers. ACC is the total efficiency of the classifier, which determines the probability of correct classification, i.e. the ratio of correct classifications to all classifications. It is expressed by the equation:

\[
ACC = \frac{TP + TN}{TP + TN + FP + FN}
\]

where [3]:
- \(TP\) – True Positive – the number of observations correctly classified to the positive class.
- \(TN\) – True Negative – the number of observations correctly classified to the negative class.
- \(FP\) – False Positive – the number of observations classified to a positive class when in fact they come from a negative class.
- \(FN\) – False Negative – the number of observations classified to the negative class when in fact they come from a positive class.

3. Results and discussion

The results of classification of training and test sets for each of the proposed models are shown in figures 1-4.

For model 1 (figure 1), which use all attributes, in case of a training set, the highest accuracy was achieved for the Random Forest classifier (99.75%). Equally high accuracy (99.03%) was obtained for the SVM algorithm. Algorithm with the lowest accuracy is Naïve Bayes algorithm (64.56%).

For test set, we can see that all the algorithms work correctly. Four of them give more than 85% accuracy. The highest accuracy was obtained using the Random Forest and SVM algorithm (99.51%), while the weakest results (64.08%) were obtained using the Naïve Bayes algorithm. Differences in accuracy in figure 1 for the training and test sets are small, which indicates a robust model.

![Figure 1](attachment:image1.png)

**Figure 1.** Comparison of accuracy (ACC) of classifiers for model 1 (all variables) for training and test sets.

Model 2 (figure 2) contains set of attributes extracted by means of feature selection – chi-square test (5 attributes: age, cement, water, superplasticizer and coarse-aggregate).
Figure 2. Comparison of accuracy (ACC) of classifiers for model 3 (variables extracted after attributes selection) for training and test sets.

As seen in figure 2 for model 2, the most accurate classifiers for the training set are Random Forest and SVM (99.75% and 99.15% accuracy, respectively). The total accuracy of the other classifiers is within 63-92%, which is also quite a good result. For the test set, Random Forest and SVM are also the most accurate classifiers (99.52%). Other classifiers also work correctly giving good matches. In principle, all of them give the test set slightly lower accuracy index than in case of the training set, which indicates a very good model.

The last model is a model which contains 3 new variables obtained by PCA principal components analysis. The classification results for this model are shown in figure 3.

Figure 3. Comparison of accuracy (ACC) of classifiers for model 4 (new variables - PCA method) for training and test sets.
The results obtained for model 3 are not very informative. Although the Random Forest algorithm achieves almost 100% accuracy for the test set, unfortunately this accuracy drops below 50% with the test set. Other classifiers also achieve very low accuracy (even below 38%). This means that for our data, the PCA principal components analysis did not give the desired results and cannot be used to build properly functioning models.

Let us now compare the correctly classified objects in individual models (figure 4). It was compared on the basis of the results obtained on the test set, as it presents the correct accuracy and usefulness of some models.

![Figure 4. Comparison of correctly classified objects in individual models (for test set).](image)

In Figure 4, we see that the most accurate algorithms for the test set turned out to be the Random Forest and SVM algorithms for model 2, i.e. the model in which the attribute reduction was applied using chi-square tests. The weakest results were obtained for model 3, the PCA principal components analysis was used, which means that it failed to work with our data. It can also be seen that model 3 (5 attributes) has basically the same accuracy as model 1, in which all attributes are taken into account. This implies the legitimacy of using attribute reduction, and the selected Chi-square test has proved to be optimal in this case.

4. Conclusion
Discovering knowledge in databases is a dynamically developing field, whose rapid development is related to the growing number of databases and the size of information collected in them. Increasingly, data mining finds its application in engineering sciences. The analysis of classification algorithms carried out shows that they can be a very important tool supporting, e.g., material classification and analysis. Detailed data analysis, proper preparation and proper classification allow achieving very accurate results.

References
[1] Aggarwal C C 2015 Data Classification Algorithms and Applications (Chapman & Hall/CRC New York) pp 1-31
[2] Bramer M 2016 Principles of Data Mining (London: Springer)
[3] Fernandez-Delgado M, Cernadas E, Barro S and Amorim D (2014) Do we need hundreds of
classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15** 3133-81  
[4] Frank E, Hall M A, Witten I A 2016 *The WEKA Workbench. Online Appendix for ”Data Mining: Practical Machine Learning Tools and Techniques”* (San Francisco: Morgan Kaufmann)  
[5] Han J, Kamber M 2006 *Data mining. Concepts and Techniques* 2nd ed (San Francisco: Elsevier)  
[6] Alaiz-Moreton H, Fernández-Robles L, Alfonso-Cendón J, Castejón-Limas M, Sánchez-González L and Pérez H 2018 Data mining techniques for the estimation of variables in health-related noisy data *Adv. Intel.Sys.* **649** 482–91  
[7] Larose D T 2005 *Discovering knowledge in data. An introduction to data mining* (New Jersey: John Wiley & Sons Inc.)  
[8] Kiranmai S A and Laxmi J A 2018 Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy *Prot. Contr. Mod. P. Sys.* **3**(29) https://doi.org/10.1186/s41601-018-0103-3  
[9] Rogalewicz M and Sika R 2016 Methodologies of Knowledge Discovery from Data and Data Mining Methods in Mechanical Engineering *Man. Prod. Eng. Rev.* **7**(4), 97–108  
[10] Yoo I, Alafaireet P and Marinov M 2012 Data mining in healthcare and biomedicine, A survey of the literature *J. Med. Sys.* **35**(4) 2431–48  
[11] Yeh I-C 1998 Modeling of strength of high performance concrete using artificial neural networks *Cement and Concrete Research* **28**(12) 1797–808  
[12] Witten I H, Frank E, Hall M A 2011 *Data Mining: Practical Machine Learning Tools and Techniques* (San Francisco: Morgan Kaufmann)  
[13] Maimon O and Rokach L (ed) 2010 *Data mining and knowledge discovery handbook* (Boston, MA: Springer)  
[14] Dardzińska A 2013 *Action Rules Mining* (Berlin: Springer)