Chemical Engineering Science 173 (2017) 131–144

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

Variable selection and training set design for particle classification using a linear and a non-linear classifier

Stefan Heisel [a,1], Tijana Kovačević [b,1], Heiko Briesen [b], Gerhard Schembecker [a], Kerstin Wohlgemuth [a,*]

[a] Laboratory of Plant and Process Design, TU Dortmund University, 44227 Dortmund, Germany
[b] Chair of Process Systems Engineering, TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85356 Freising, Germany

HIGHLIGHTS

- A procedure for classifier generation for particle classification is proposed.
- Particles are discriminated into single crystals, agglomerates, and gas bubbles.
- Discriminant factorial analysis is compared to artificial neural networks.
- Artificial neural networks need fewer variables that describe the particles.

ARTICLE INFO

Article history:
Received 24 February 2017
Received in revised form 3 July 2017
Accepted 16 July 2017
Available online 18 July 2017

Keywords:
Discriminant factorial analysis
Artificial neural networks
Crystallization
Image analysis
Agglomeration
Bubbles
Training set

ABSTRACT

While particulate products are often characterized by their median diameter or the width of the particle size distribution, information is rarely given about the agglomeration degree of the product. To obtain this information, a tool combining image analysis and discriminant factorial analysis (DFA) was introduced in previous works. The accuracy of that method depended on the number of image descriptors selected, i.e. measurements describing each particle: few image descriptors resulted in rather poor classification while too many lead to an overfitting of the data. The aim of this study is twofold: First, we want to compare the classification accuracy of artificial neural networks (ANN) and DFA which, contrary to ANN, forms linear classifiers. Second, we want to provide an easy-to-implement procedure for generating particle classifiers. We used a qualitative measure called *Proportional Similarity* to test whether a subset selection of image descriptors was necessary to avoid an overfitting. The influence of the training set size was investigated as well as the transferability of the classifier on data obtained under different experimental conditions. The chemical systems used were L-alanine/water and adipic acid/water and the classes considered were single crystals, agglomerates, and gas bubbles. The results show that an ANN classifier provides higher accuracy and is more effective when only few image descriptors are available while DFA is simpler to create. Moreover, we show good transferability of classifiers trained on data of different experimental conditions. Based on our results, we provide guidelines for classification of particulate systems.

1. Introduction

Due to the increasing demand of specific product characteristics, optimization of the process chain for production of particulate products is of high interest in the chemical, pharmaceutical and food industry. Next to bulk densities and the specific surface area, the particle size distribution is a common measure for ensuring the correct product specifications. In this contribution we focus on agglomerates, consisting of solidly bound single particles, which may have an important effect on the final product. We use the term agglomeration degree (*Ag*), which is defined as the ratio of the number of agglomerates and the overall number of particles, to specify the extent of agglomeration of the product. However, it is hardly possible to infer the information on *Ag* only from the particle size distribution or other commonly used characteristics without considering the shape of the particles. In this paper, we focus on crystal agglomeration which affects purity, flowability (Mullin, 2001) and filterability (Beckmann, 2013) of the crystalline product batch.

The core of studying shape related properties is typically performed via image processing. Particle characterization methods based on two-dimensional (2D) (Schorsch et al., 2014; Borchert

* Corresponding author.
E-mail address: kerstin.wohlgemuth@bci.tu-dortmund.de (K. Wohlgemuth).
[1] Shared lead authorship.

http://dx.doi.org/10.1016/j.ces.2017.07.030
0009-2509/

et al., 2014; Zhang et al., 2015) and three-dimensional (3D) (Singh et al., 2012; Kovačević et al., 2014, 2016) image processing have gained importance in recent years, enabling a sophisticated measurement of size and shape. The information obtained from 2D imaging techniques can also be employed for characterizing the agglomeration degree. To this end, pattern recognition methods are used to classify particles into single crystals and agglomerates. Several researchers applied discriminant factorial analysis (DFA) on off-line (Faria et al., 2003; Yu et al., 2007; Terdenge et al., 2015; Terdenge and Wohlgemuth, 2016) and on-line (Ferreira et al., 2011) images of organic crystals. Ochsenbein et al. (2015) used a non-linear method called support vector machine and a stereoscopic imaging setup to discriminate between single crystals and agglomerates of needle-like β L-glutamic acid. Ålander et al. (2004), Ålander and Rasmuson (2005) investigated the influence of solvent on agglomeration of paracetamol based on principal component analysis. The classification was done by correlating the ratio of principal components to the number of primary particles in an agglomerate. Similar problem statements also occur when considering granulated particles in food industry, which motivated Ros et al. (1995) to investigate artificial neural networks (ANNs) for detecting whether the imaged objects overlap. Apart from measuring the agglomeration degree, pattern recognition tools can be used to classify crystals according to their shape (Bernard-Michel et al., 1997; Hundal et al., 1997) or polymorphic form (Zhang et al., 2016; Huo et al., 2016).

DFA and ANN are pattern recognition methods and can be applied to a wide range of problems where objects or patterns need to be sorted into predefined classes. To do so, measurements need to be performed on the objects that are to be classified, resulting in variables that describe each of the objects' attributes. DFA and ANN use a set of manually classified objects – the training set – along with the corresponding variables to generate a set of mathematical operations – the classifier – that assigns the objects to the predefined classes. The accuracy of the classifier is measured on a separate test set, containing objects different from the training set. Finally, the classifier obtained is used to classify previously unseen objects, thus automatizing the classification process. The DFA classifier is also called the discriminant function and it forms linear boundaries between classes. It is a parametric method as the underlying algorithm assumes that all variables are normally distributed for all classes and finds the parameters of these normal distributions. Contrary to that, ANN is a nonparametric method since no parametric representation of variables' probability distributions is assumed. The classifier is an interconnected network. Each node (neuron) performs some function over the inputs and the output neurons give the membership probability for each class. The boundaries obtained between classes are non-linear.

Before discussing the classifier design for measuring the agglomeration degree, we review the general properties of DFA and ANN classifiers. The classification accuracy depends on the classification method chosen, the number of objects in the training set (Hamamoto et al., 1996; Hand, 1981; Friedman, 1989), the number and quality of the variables selected (Hand, 1981; Raudys and Young, 2004) and the relative class sizes in the training set (Foody et al., 1995). In case of ANN, the results also depend on the number of neurons in the network (Hamamoto et al., 1996). In many applications, the size of the training set is limited by the costly measurement procedure (Gupta et al., 1999), whereas a large number of variables can be measured simultaneously for each object (Sima and Dougherty, 2008). The resulting small but high-dimensional training set can lead to the so-called "peaking phenomenon" (Raudys and Young, 2004; Sima and Dougherty, 2008; Raudys and Jain, 1991): For an increasing number of variables the classification results for a test set improve, only to worsen after a certain number of variables is exceeded. This phenomenon is explained by the fact that "the useless features [variables] produce only parameter estimation errors" (Hamamoto et al., 1996), leading to an overfitting of the training set. It can be avoided by either increasing the training set size or selecting an appropriate subset instead of using all variables available. When using DFA, a lot of effort is invested into the time-consuming subset selection, both in case of crystal classification (Faria et al., 2003; Terdenge et al., 2015) and in other applications (Pacheco et al., 2006). ANN classifiers appear to be more robust regarding the "peaking phenomenon" as they are capable of neglecting the useless variables (Hamamoto et al., 1996).

The aim of this study is twofold. First, we want to compare the accuracy of DFA and ANN classifiers with respect to measuring the agglomeration degree of a crystalline product batch. Second, we want to develop a procedure for selecting the necessary variables and designing training sets for particle classification without using the brute-force search for finding the best variable subset in $2^{\text{number of variables}} - 1$ possible combinations. This procedure shall be easy to implement and consider all important parameters for creating a classifier that delivers classification results with an accuracy of at least 90%. To do so, we investigated the training set size and the selection of variables, where we suppose that a suitable variable subset can be determined by using an easy to implement measure called *Proportional Similarity* (*PS*). *PS*, according to Vegelius et al. (1986) first introduced by Renkonen (1938), is one of the most useful measures to quantify the similarity between two classes that are distributed over a certain variable (Vegelius et al., 1986). Moreover, the influence of the underlying experimental conditions on the transferability of classifiers between samples is investigated. The classes considered in the analysis are single crystals, agglomerates, and gas bubbles. During a crystallization experiment, different agglomeration degrees may occur. If the objects for the training set are sampled randomly, their relative amount will depend on the agglomeration degree in the underlying experiment, as well as on the amount of bubbles. This can allow the classifier to "learn" some classes better than others or even act as a set of a priori class membership probabilities in case of ANN (Foody et al., 1995). In order to avoid this, we only consider training sets containing the same number of objects in each class. In our work, the variables are the so-called image descriptors which describe the objects' size and shape. Therefore, in the following we use the term image descriptors instead of variables. To achieve our aims mentioned above, we test the validity of the following hypotheses:

1. ANN outperforms DFA
2. *PS* can be used for image descriptor selection, even for distinguishing three classes
3. Larger training sets improve the classification accuracy
4. A classifier shows poor transferability to samples of different experimental conditions

To prove whether these hypotheses hold for different crystal morphologies, we test two different chemical systems: L-alanine/water which forms medium agglomerated bipyramidal structures and adipic acid/water which forms highly agglomerated platelets. We consider three different experiments of each of the systems, obtained under different process parameters. The experimental products differ in width of crystal size distribution (CSD) as well as overall agglomeration degree.

The paper is structured as follows: In Section 2 the experimental setup as well as the image analysis steps to obtain the image descriptors are presented. Furthermore, the composition of the training sets used is described in detail. In Section 3 we show and discuss the results concerning *PS*, DFA, ANN, and the comparison of both classification methods. Finally, in Section 4 a conclusion is presented with guidelines concerning the creation of a classifier.