Chemometrics and Intelligent Laboratory Systems 149 (2015) 90–96

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Pattern recognition in chemometrics

Richard G. Brereton

School of Chemistry, Cantocks Close, Bristol BS8 1TS, United Kingdom

ARTICLE INFO

Article history:
Received 26 March 2015
Received in revised form 16 June 2015
Accepted 18 June 2015
Available online 26 June 2015

Keywords:
Pattern recognition
Partial least squares discriminant analysis
SIMCA
Linear discriminant analysis
Historic review
Support vector machines

ABSTRACT

The origins of chemometrics within chemical pattern recognition of the 1960s and 1970s are described. Trends subsequent to that era have reduced the input of pattern recognition within mainstream chemometrics, with a few approaches such as PLS-DA and SIMCA becoming dominant. Meanwhile vibrant and ever expanding literature has developed within machine learning and applied statistics which has hardly touched the chemometric community. Within the wider scientific community, chemometric originated pattern recognition techniques such as PLS-DA have been widely adopted largely due to the existence of widespread packages, but are widely misunderstood and sometimes misapplied.

1. Introduction

There is no universal definition of Pattern Recognition (PR), the definitions being slightly different within disciplines such as statistics, computing, engineering and so on. PR has a strong overlap with machine learning, data mining and classification.

The first recognisable corpus of knowledge began in the late 1960s. The journal *Pattern Recognition* [1] pulled together what might be considered quite a wide variety of techniques in what was then the new discipline of computer science, involving feature selection, character recognition, classification etc. As time moved on, pattern recognition has increasingly been concerned with discrimination (or classification). In 2002, Webb states [2] "*Statistical pattern recognition is a term used to cover all stages of an investigation from problem formulation and data collection through to discrimination and classification, assessment of results and interpretation.*" Jain et al. [3] state "*The primary goal of pattern recognition is supervised or unsupervised classification.*" Duda et al. [4] define pattern recognition as "*the act of taking in raw data and taking action based on the category of the pattern*". The on-line dictionary of computing [5] states "*Pattern recognition aims to classify data (patterns) based on either a priori knowledge or on statistical information extracted from the patterns.*" Fukanaga [6] states "*pattern recognition, or decision-making in a broader sense, may be considered as a problem of estimating density functions in a high-dimensional space and dividing the space into the regions of categories or classes.*" Theodoridis and Koutroumbas [7] state "*Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes.*"

In common with most modern definitions of PR, it involves primarily classification to assign objects into groups or classes or categories. The earlier pioneers of pattern recognition broadened the definition to almost any computational approach used to determine patterns or relationships between objects, e.g. handwriting analysis or facial recognition, but we will stick to the more focussed definitions that involve some form of classification. A question might arise whether the title "pattern recognition" should just be changed to "classification". There is no completely agreed answer, however classification should be regarded as an aspect of pattern recognition. Whereas we might be able to ask to classify a fruit into an apple or an orange, PR will do more than that. It may determine how well we can do this classification and so validate the model. It may determine how many classes and also whether there are outliers. It may determine what features are best at distinguishing these groups. So although classification algorithms are at the centre of most pattern recognition methods, they are not exclusive.

Classification methods can be divided into two types. Supervised approaches attempt to divide objects into groups according to their characteristics using a training set, that is objects that are labelled into predefined categories. Unsupervised approaches attempt to divide up dataspace into groups without any predefined training set. Most pattern recognition involves supervised learning, and will be the focus of this paper. Unsupervised approaches can involve methods such as cluster analysis.

Methods for PR can be further divided into purely statistical methods and machine learning, although nowadays there is no sharp distinction. However classical methods such as Linear Discriminant Analysis, first developed by Fisher in the 1930s [8], involve methods that are considered to be primarily statistical in basis. These often do not involve

E-mail address: r.g.brereton@bris.ac.uk.

http://dx.doi.org/10.1016/j.chemolab.2015.06.012
0169-7439/

elaborate computations and the answer can formulated by simple mathematical equations. Methods involving machine learning were developed somewhat later and can involve neural networks [9] and Support Vector Machines [10]. Of course modern approaches for machine learning often do have statistics, especially Bayesian, buried within algorithms but are normally computationally intense, and in some cases do not have a reproducible solution.

As we will see, although in the beginning PR had a major role to play in chemometrics, this dominance reduced substantially as time went on.

2. Rise and fall of pattern recognition in chemometrics

2.1. The growth and origins

In the 1960s to 1980s PR was considered a major growth point in chemometrics.

In the US, the group of Isenhour, Jurs and Kowalski published a significant number of papers on what we would call PR often including linear learning machines, examples of which are [11–13]. Kowalski effectively took over the laboratory and in the early days continued to publish work primarily on PR [14] viewing most of the work in Washington over several years as chemical pattern recognition. In 1975, although Kowalski had started to adopt the word chemometrics, he still viewed pattern recognition as the key stating "*Computer Pattern Recognition Methods extend the capacity of Human Pattern Recognition Methods*" [15]. Many of the early papers In this group related to the development of machine learning techniques, but broadened and changed in the mid 1970s with the introduction to multivariate statistics. In 1975, following on from an ACS Symposium on Chemical Applications of Pattern Recognition, Kowalski published an article on Chemometrics: Views and Propositions [16]. Kowalski and coworkers continued a focus on PR: a survey of 7 publications in 1980 by Web of Science, reveals 4 of them having titles and therefore focus on "pattern recognition". Prior to a series of reviews starting in 1980 entitled "Chemometrics" the last review on Statistical and Mathematical Methods in Analytical Chemistry [17] states under pattern recognition "*In fact, the most frequently appearing papers on data processing techniques have been in this area. Probably the simplest statement of the pattern recognition process is the transformation of patterns from measurement space into classification space.*" and this topic occupies around two and a half out of six pages of text, or close to half the article.

Wold is credited with first introducing the word chemometrics to the literature. Wold's earliest work was in areas such as curve-fitting and kinetics, and he only started publishing about multivariate methods in 1974 [18]. He described these methods as pattern recognition [19]. Wold's earliest contributions to what we would now call chemometrics (as opposed to kinetics and curve-fitting) was to develop the method of SIMCA, which is usually considered a one class classifier using disjoint principal component analysis [20]. Many of the early papers on SIMCA and related methods include the words pattern recognition in the title [21] focussed still on classification, but with elaborations such as outlier detection and determining important variables or features as part of overall strategies and regression where it enhances classification models by providing additional information, for example we may ask whether a compound is toxic, and then how toxic it is. Of 45 papers published by Wold and colleagues and recorded in Web of Science between 1975 and 1980, 20 (or 44%) have the word "pattern recognition" in the title, far more than chemometrics. Of those left out many are about other diverse subjects such as kinetics, Hammett relationships and splines. Hence although the label was chemometrics, in practice a significant amount of early work by Wold and his collaborators in Sweden was viewed as PR, and most of the early advances involved SIMCA which was developed as an aid to classification.

Multivariate statisticians, meanwhile, had been working on PR techniques since the 1930s, catalysed by Fisher's classical paper on the discrimination of iris data by their physical characteristics [8]. Numerous books, conferences, and papers had been published over the decades during the early years of chemometrics. However very few such studies have been reported in the chemometrics literature. Friedman is one of the few classically recognised statisticians that published in chemometrics [22]. It is interesting that this paper in 1989 whereas over 10 years behind several of what are considered the classic articles in chemical pattern recognition cites only 8 published references. In addition many of the classic texts and workers in statistical pattern recognition some of which are cited in the introduction have never published in, and in many cases are largely unknown in, the chemometrics literature. The very modest input of statisticians to the chemometrics community contrasts completely with the very active level of publications in the general statistical community, a few key texts being referenced in the introduction, and suggests quite strong divergence between the applied statistics and chemometrics community.

A fourth strand emerging at the time was in the continental analytical chemistry community. The focus was much more diverse and early chemometrics was by no means dominated by this theme. Strouf [23] and Varmuza [24] wrote early texts and worked In this area. Forina and the group in Genova developed Parvus [25] which was described by Massart as "*a package for pattern recognition, which …. must be the most complete such package written directly for microcomputers. It contains almost all display methods and supervised learning methods currently used by analytical chemometrists.*" Massart developed the UNEQ method [26] which is a one class classification technique with some similarities to SIMCA. Unsupervised learning, primarily cluster analysis, was also of interest to continental chemometricians [27]. However this strand of chemometric thinking started around ten years after the work centred around Kowalski and Wold, and several decades after mainstream statisticians. The large number of approaches developed whereas probably quite worthy and in some cases more appropriate and show no less promise theoretically, have not gained great currency. As we will see below, the SIMCA and the later described PLS-DA (Partial Least Squares Discriminant Analysis) approaches are far more widespread.

In conclusion, many of the pioneers of chemometrics actually worked in the area of multivariate pattern recognition, starting from the mid 1960s up to the 1980s. Applied statisticians briefly entered the fray but largely moved away into their own circles. A small strand developed in continental chemometrics but several years later.

We will explore how PR has laid low over the past few decades as a development area for chemometrics.

2.2. The decline

Despite the origin of chemometrics being in chemical pattern recognition, the early growth years show relatively limited expansion of interest in this area. As chemometrics developed as a named discipline, many other people climbed onto the bandwagon. In some ways this was inevitable as new researchers from new perspectives joined the community which then became a more significant force. Although there were many potential applications of PR in chemistry, the capacity, or indeed motivation, of early researchers to obtain large datasets was limited. The early description of the PARVUS package limits data to 10 variables, something that would be considered a major limitation in the 2010s where huge numbers of variables can be routinely obtained, from say mass spectrometry or NMR or NIR spectroscopy. Many of the early chemical pattern recognition papers involved very few samples, 15 in a group (which would then need to be divided into training and test sets) would often be regarded as quite a substantial dataset. Although challenging data was available, because the size was usually limited, manual methods of interpretation often offered acceptable and feasible alternatives to the chemist.

As time went on, a plethora of methods became incorporated into the chemometrician's toolkit. These included methods for multivariate curve resolution, called by some factor analysis and others alternating least squares (where appropriate) or deconvolution, pioneered by