Pattern Recognition Letters 64 (2015) 53–62

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

What are the true clusters? ☆

Christian Hennig*

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom

ARTICLE INFO

Article history:
Available online 29 April 2015

Keywords:
Constructivism
Active scientific realism
Natural kinds
Categorization
Mixture models
Variable selection

ABSTRACT

Constructivist philosophy and Hasok Chang's active scientific realism are used to argue that the idea of "truth" in cluster analysis depends on the context and the clustering aims. Different characteristics of clusterings are required in different situations. Researchers should be explicit about on what requirements and what idea of "true clusters" their research is based, because clustering becomes scientific not through uniqueness but through transparent and open communication. The idea of "natural kinds" is a human construct, but it highlights the human experience that the reality outside the observer's control seems to make certain distinctions between categories inevitable. Various desirable characteristics of clusterings and various approaches to define a context-dependent truth are listed, and I discuss what impact these ideas can have on the comparison of clustering methods, and the choice of a clustering methods and related decisions in practice.

© 2015 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cluster analysis is about finding groups in a set of objects. Cluster analysis is used in many different areas with many different aims (see Section 3 for examples). Researchers who apply cluster analysis in practice often want to know whether the clusters that they find are truly meaningful in the sense that they correspond to a real underlying grouping. Researchers in the field of cluster analysis are interested in whether and which methods are better at finding the true clusters correctly. In most cluster analysis literature, however, explanations of what "true" or "real" clusters are, are rather hand-waving. It is widely acknowledged that there is no agreed definition of what a cluster is, and in the majority of papers in which new cluster analysis methods are proposed, the authors do not give a general and formal definition of what the "true clusters" are that their method is supposed to find.

The aim of this paper is to offer a philosophically informed attitude toward the problem of choosing, assessing and interpreting cluster analysis methods and clusterings. Section 2 gives an overview of thoughts in philosophy and cognitive science regarding clustering and categorization. Afterward the paper turns to considerations and implications that are directly related to the theory and practice of data-based cluster analysis.

The groups that cluster analysis sets out to find are characterized by data that can take various forms such as values of variables, dissimilarities or weighted edges in a graph. The groups may form a partition of the object set, but they may also be overlapping or non-exhaustive. Group memberships may be crisp or fuzzy. Some of the discussion here was written with crisp partitions in mind, some apply to Euclidean space or a given dissimilarity measure, but most thoughts are more general.

There is a good reason why there is no generally accepted unique definition of true clusters. In different applications, cluster analysis is used with different aims, and the researchers have different ideas of what should make the objects belong together that are in the same cluster. The term "cluster" does not mean the same to all researchers in all situations. This is acknowledged in general overviews and books about cluster analysis, but seems to be ignored by many authors of specialist work who try to convince readers that a certain method is best for finding the "true/natural/real" clusters. Even where it is acknowledged, this often takes the form of a "general health warning", and consequences regarding the selection and comparison of methods and the interpretation of results are rarely spelled out. Is it possible to escape the alternative to either make the hardly justifiable assumption that there is a unique "true/natural/real" clustering against which the quality of cluster analysis methods can be objectively assessed, or to think that cluster analysis is somehow arbitrary and "more of an art than a science" [1]?

My perspective is that of a statistician with expertise in cluster analysis and a strong interest in the philosophical background of statistics and data analysis. A key idea of this paper is that, given that it depends on the context and clustering aim what a "good" clustering is, researchers need to characterize what kind of clusters are required for a given real clustering problem, and what

☆ This paper has been recommended for acceptance by Marcello Pelillo.

* Tel.: +44 020 7679 1698.
E-mail address: chrish@stats.ucl.ac.uk, c.hennig@ucl.ac.uk

http://dx.doi.org/10.1016/j.patrec.2015.04.009
0167-8655/© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

kind of clusters the different clustering methods are good at finding, or in other words, what problem-specific "truth" researchers are interested in. Similar ideas have recently been discussed in [2] and [1]. The present paper can be seen as contributing to the research program sketched in those papers, but also as enrichening their perspective by adding further philosophical and statistical considerations.

In Section 2 I will sketch the philosophical basis of the present paper, which complements constructivism with Hasok Chang's pluralist active scientific realism, and I will discuss the concepts of "natural kinds" and "categorization". Section 3 lists and discusses various context-dependent clustering aims. Section 4 is about how "true" clusters could be defined in statistical or data analytic terms so that they can be used for comparing and assessing different clustering methods. Section 5 discusses some practical consequences, particularly regarding choice and comparison of cluster analysis methods, and rationales for certain methodological decisions such as dimension reduction.

2. Philosophical background

2.1. Constructivism and science

In the present paper I focus on the question what clusters are "true" and/or "real". Truth and reality, and to what extent they can be observed, are controversial issues in philosophy. My starting point in this respect is my constructivist philosophy of mathematical modeling as outlined in [3], which is connected to radical constructivism [4] and social constructionism [5]. Radical constructivism is based on the idea that the perception and world-view of human beings can be interpreted as a construction by the body and the brain of the individual, which is seen as a self-organizing system. Social constructionism focuses on the construction of a common world-view of social systems by means of communication. "Construction" refers to the activity of the body, the brain, and communicative activity within social systems, setting up perceptions and world-views. Construction is largely unconscious or semi-conscious, and is not arbitrary but subject to constraints. It is not claimed that individuals or social systems are free to construct any arbitrary perception or world-view. Experience tells us that perception is rather severely constrained and shaped by what we perceive to be a reality outside of ourselves.

I distinguish observer-independent reality, personal reality and social reality. The observer-independent reality is only accessible to humans by observation, which means that there is no way to make sure which of its features are really observer-independent, but it is usually perceived as the source of constraints for personal and social constructs. The perceptions of individuals, together with their thoughts and feelings, make up their personal reality. Part of most personal and social realities is the belief that much personal perception represents or reflects the observer-independent reality. This belief is normally based on the experience of consistency between different sensory perceptions, at different times and from different positions, and on the confirmation of the existence of many of the perceived items by communication with others. It is therefore the result of active accommodation of perceptions.

Social reality is made up by communication between individuals. It is carried by social systems, which may overlap and may partly lack clear borderlines, although some social systems such as formal mathematics are rather clearly delimited. Personal and social realities influence each other. According to the point of view taken here, science is a social attempt to construct a consensual and stable view of the world, which can be shared by everyone and is open to criticism and scrutiny in free exchange. In this sense, science aims at a view that is as independent as possible of the individual observer, and is therefore connected to a traditional realist view, according to which science aims at finding out the truth about observer-independent reality. But constructivists are pessimistic regarding an observer-independent access to reality, and assess the success of science based on stability, agreement and pragmatic use instead of referring to objective truth. A scientific world-view with which constructivists can agree needs to acknowledge the existence and legitimacy of diverse personal and social realities and is therefore inherently pluralist. A tension between a drive for unification and general agreement and a necessity to allow space for diverse realities in order to allow for criticism and creative progress is an essential implication of the scientific idea. Central tools of science are mathematics, which aims at setting up and exploring concepts that are clear and well defined independently of the different personal and social points of view and at statements about which absolute agreement is possible, and measurement, which unifies observations of reality in a way that they can be processed by mathematical means.

Constructivism is often accused of denying the existence of the observer-independent reality altogether by calling it "a construct", but actually, being as stable and ubiquitous a construct as the observer-independent reality seems to be in most personal and social realities, it is as real as anything can get in constructivism.

2.2. Active scientific realism

Although constructivism is often interpreted as anti-realist, I complement my constructivist view here by the "active scientific realism" introduced by Hasok Chang [6]. In the abstract of his Chapter 4, Chang writes: *"I take reality as whatever is not subject to ones will, and knowledge as an ability to act without being frustrated by resistance from reality. This perspective allows an optimistic rendition of the pessimistic induction, which celebrates the fact that we can be successful in science without even knowing the truth. The standard realist argument from success to truth is shown to be ill-defined and flawed. I also reconsider what it means for science to be "mature", and identify humility rather than hubris as the proper basis of maturity. The active realist ideal is not truth or certainty, but a continual and pluralistic pursuit of knowledge."* Chang's use of the term "reality" refers to what is vital for the success of the scientific idea, namely to confront scientific work continually with the observed realities that individuals and social systems experience as outside their control. In agreement with my constructivist view, active scientific realism values a plurality of perspectives. The term "truth" is constructivist used in both Chang [6] and the constructivist literature as a relative concept "internal to systems of practice". For example, within the mathematical formal system, "truth" is a rather unproblematic concept due to the clear rules by which it can be ensured, whereas the truth-value of the statement "the German Democratic Republic was a democracy" depends on which characteristics of a political system are taken as essential for being a democracy, which differs between social systems.

The emphasis of the strong role of communication and language is an aspect that constructivism adds to active scientific realism. In this respect I follow Fleck [7], a pioneer work regarding the role of communication and social systems ("thought collectives") for scientific knowledge. Fleck showed how scientific facts are shaped by the specific way how collectives of scientists conceptualize their field.

2.3. Natural kinds

"Natural kinds" in philosophy refer to the idea that there are some "naturally" separated classes in observer-independent reality, which, for traditional realists, correspond to "true clusters". For example, biological species and chemical elements are considered as candidates for being natural kinds [8]. There is much controversy about what constitutes natural kinds (e.g., common properties, behaving homogeneously according to natural laws). The concept runs counter to the constructivist view that what is perceived as "kinds" is constructed by human activity and language and depends on the conditions of observation and practice of living of the observers. For such reasons, for

Download English Version:

https://daneshyari.com/en/article/6941189

Download Persian Version:

https://daneshyari.com/article/6941189

Daneshyari.com