On the futility of attempts to formalize clustering within the conventional mathematical framework

Lev Goldfarb[*]
Faculty of Computer Science
University of New Brunswick
Fredericton, NB
goldfarb@unb.ca

In the workshop announcement we read: "Despite this large number of algorithms and applications, the goal of clustering and its proper interpretation remains fuzzy and vague". In connection with this, the main point I want to address is that the above "fuzziness" and "vagueness" is a reflection of a situation which cannot be rectified within the conventional mathematical setting, e.g. that of a metric space or a normed vector space. **The main reason is this**: *within the traditional mathematical paradigm*, the central concepts relevant to clustering cannot be adequately addressed simply because the most basic, underlying *concept of class*—i.e. the concept on which all of the former concepts depend, either explicitly or implicitly—*cannot be adequately formalized*.[1]

In turn, the reason the concept of class cannot be formalized has to do with the impossibility of introducing a meaningful *concept of class representation* within the conventional mathematical framework. The latter observation has to do with the very structure of modern mathematical language in general, and the axiomatic structure of classical spaces in particular: all classical mathematical structures are introduced via several kinds of axiomatically specified *relations* defined over some *sets* of objects (including Cartesian products). Moreover, since the *elements of an underlying set are not specified as having any internal structure*, they may acquire *some non-hierarchical structure* only indirectly, if at all, via the specified relations: e.g. in a vector space, a vector can be represented as a sum of several other vectors.

Before addressing the main argument, it is interesting to note (and useful to keep in mind) that the founder of set theory, George Cantor, thought of the concept of set as being quite close to that of a class, i.e. he thought of it as "a multitude that *can be thought of* as one": after all, a class *is* such a "multitude". To some extent this confusion is not surprising, since it appears that the basic "unit" of biological information processing is a class, and not a set. Thus, when developing a representational formalism, one should approach the concept of class as being at least as fundamental as that of a set, requiring, however, a radically new formal language for its explication.

More to the point, there are strong reasons to believe that the concept of class cannot be properly formalized without simultaneously addressing the twin *concept of class representation*, which is supposed to specify a *mechanism by means of which the members of the*

[*] http://www.cs.unb.ca/∼goldfarb

[1] I'm saying this despite the fact that, twenty seven years ago at the University of Waterloo, I almost wrote my doctoral thesis on clustering.

class are "generated". The reason why the most popular machine learning formalisms—developed within the classical vector space representational paradigm—have not addressed the concept of class representation is directly related to the above observation: this concept *simply cannot* be adequately formalized within that paradigm. As a result, in machine learning and pattern recognition, the concept of class representation is avoided altogether by replacing it with the absolutely non-revealing, or structurally non-committal, i.e. not carrying any structural load, concept of indicator function.

Why cannot the concept of class representation be adequately formalized within the traditional mathematical framework? Here is a very brief outline of the argument. From an applied point of view, we are compelled to treat all objects in the Universe (including man-made objects) as having some hierarchical structure that has evolved gradually and concomitantly with the structure of their classes. Hence, an adequate representational formalism must offer, up front, a *dynamic formal structure* for capturing the evolving hierarchical structure of both objects and classes. I wish to draw your attention to precisely these two absolutely critical features: if, in a chosen representational formalism, the object or class representations cannot be *naturally* treated as being *both hierarchical and dynamic*, the battle for class representation has already been lost. I suggest that such formal capabilities must be "visible" at the outset, i.e. they cannot be "introduced" later. Specifically, if the basic entities in a formalism—e.g. the elements of a set in modern mathematical structures—are treated as non-structured, the latter two features cannot be adequately introduced into the formalism: in this case, the hierarchical structure can only be *postulated*, in which case this structure is fixed and cannot be modified dynamically.

Consider, for example, the *vector space as a representational formalism*. Its main hierarchical structure captures the relationship among various (linear) subspaces but has very little to do with the structure of "classes" as they emerge from various applications. Partly as a result of this, we have—from a formal point of view—a very unsatisfactory situation. On the one hand, the *definition of a class* is structurally vacuous: it is defined via a *structurally non-committal* indicator function. On the other hand, at the end of learning, we end up with a (non-generative form of) *class "representation"* specified via piecewise linear or non-linear surfaces. Consequently, we have no meaningful representation of either a (non-training) class object or of the class itself[2]: to gain such knowledge, the representational formalism must commit, up front, to a *structurally meaningful form of class representation*, instead of a structurally vacuous indicator function.

For the last six years, we have been developing a radically new representational formalism [1] (for more on class representation, see also [2]), developed specifically with the goals of inductive learning in mind. At the workshop, I will outline the main features of this new formalism.

References

[1] Lev Goldfarb, David Gay, and Oleg Golubitsky, What is a structural representation? Fourth variation, Faculty of Computer Science, U.N.B., Technical Report TR05-174, July 2005 (also submitted to Pattern Recognition). `http://www.cs.unb.ca/~goldfarb/ets4/ETS4.pdf`

[2] Lev Goldfarb and David Gay, Class representation as the only source of inductive transfer, submitted to NIPS '05 workshop on Inductive Transfer. `http://www.cs.unb.ca/~goldfarb/nips2005/transfer.pdf`

[2] Even in the case of a piecewise linear class specification—which, from the representational point of view, is less arbitrary than a non-linear specification (after all, the underlying space has linear structure)—the learned class description is not only "unstable" but, most importantly, it offers us nothing from the point of view of "inductive transfer" [2].