

Clustering: Knowledge From Understanding

Joey Kippen

December 16, 2009

Contents

1	Introduction	1
2	Motivation	1
3	Methods	2
4	The Papers	2
4.1	A Review	2
4.2	The Algorithms	3
4.3	Unification	3
5	Implications	3
6	Conclusions	4

1 Introduction

In machine learning, there are several different ways of using data. Data can be used to predict values, classify data, or calculate patterns. Pattern recognition is the ability of a computer to discern patterns in data. A sub area of pattern recognition is called clustering. This area of machine learning is used to create categories in data based on commonalities based on certain criteria for some meaningful purpose. Since this is the area of interest the following sections are organized in this way. The next section, Section 2, explains the motivations that directed me to this area. Section 3 describes how I intended to expand my knowledge in this matter. After that, Section 4 summarizes the information that was found on the topic followed by an in depth look at the implications of those papers in Section 5. Lastly, in Section 6, some general conclusions about the area of clustering and generalizations are made.

2 Motivation

Generalization of data has been an interest of mine in machine learning for some time now. However, classes in machine learning that were available were so generalized themselves that pattern recognition was only an area to overview and delve into. Therefore, when I was allowed to contemplate what assignment would be interesting, I decided to investigate classification in data, having the computer generate classes instead of training the computer to match data to predetermined classes. Dr. Anderson stated that the activity I had just described was called clustering. If a computer could be taught to categorize data, then computers could be programmed to run experiments on their own and make conclusions independent of human researchers and therefore work in tandem with humans to expand knowledge further faster.

3 Methods

In an attempt to learn more in depth material on the matter, the Journal of Machine Learning Research was consulted and I began reading articles that covered the topic. The initial set of articles available was very specific and I quickly realized that all the classes taken before this time had never gone into much detail about clustering. I only knew of it in theory. So, to gain understanding of what possibilities and problems clustering had, I would need to cultivate a greater understanding about where research had been and seemed to be going. The most concise way of doing this would be to find three articles: one to summarize where research in clustering was at a certain point in time, why it was in that state and the possible faults it had, and either find explanations as to why the limitations were there, so as to work around such speculations, or what research needed to be accomplished to find such expansions to clustering.

4 The Papers

4.1 A Review

Several articles were considered for the need to accumulate a relatively large amount of information on this area of machine learning but three were chosen for very specific reasons. The first article, Data Clustering: A Review, was selected to learn more about the state of clustering as a whole. This article attempts to capture the magnitude of the clustering problem in a single paper, much like this one. It starts with definitions and descriptions about clustering that differentiate it from data analysis and gives a history on the context of clustering as well. It defines common terms such as pattern representation, feature selection and extraction, pattern proximity, cluster validity, and others. The article also stresses that human expertise in the area that the data originates allows for optimization of clustering algorithms based on the knowledge of where cluster cores are more likely to be and how many cluster to expect as well as other insights that the algorithm can use. The article moves on to show common notation in clustering and give specific definitions that will be used throughout the paper. Next it introduces how data is prepared and the different forms it can take: qualitative and quantitative. The next section of this article discusses how the similarities in the data are exposed. This is usually done by a distance metric, Euclidean (square root of the summation of squared differences of the vectors) and Mahalanobis (the difference of the vectors times their covariance matrix times the transpose of the difference vector) [2]. It can also be done by comparing to a mutual neighbor or by pre-defined concepts, areas of interest.

The next section of Jains article contains the algorithms and techniques. It begins with the techniques such as starting with a single cluster and breaking it up or starting with each point as a cluster and merging them (agglomerative vs. divisive), considering all features or one at a time (monothetic vs. polythetic), keeping all clusters separate or allowing for overlap (hard vs. fuzzy), deterministic vs. stochastic, and incremental vs. non-incremental [2]. After this a number of algorithms are presented. Each is a combination of the considerations listed in the beginning of this paragraph. They range from straight distance comparisons to hierarchical models to neural networks and genetic algorithms. For brevity's sake they will not be discussed in the same depth in this paper. In Jains article, each algorithm was discussed and strengths and weaknesses were presented. This section was concluded with a comparison of the algorithms. In total, a common theme was found with nearly all of them in that many of them failed to produce adequate results in large samples or high dimensionality. This was found because the computational complexity of the calculations and/or the complexities of the parameters needed to be adjusted for the context. In fact, only the k-means algorithm and a specific neural net equivalent, the Kohonen net, have even been tested on large data sets and show promise [2]. Larger data sets are considered on the whole in this paper as a challenge that could be met in several ways, most commonly, by using a combination of techniques for initial seeding or partitioning and clustering parts of the data at a time then considering those clusters in totality at the end of the cluster calculations.

The paper ends with a discussion of the types of application that clustering is used for. The major categories of note were image segmentation, object and character recognition, information retrieval, and data mining. These application range in the size of the data set from small to large, as listed earlier. They range in scope from gathering information from bar codes to geological analyses of landforms to internet wide data extraction [2]. Each area has algorithm types that do better on one task than another. The paper

also outlines the difficulties that even the best algorithms have with the data. This means that for each problem there are many algorithms to address specific difficulties within each problem set. This leads to many different algorithms for just one problem, but there are other reasons for the multitude of algorithms as well.

4.2 The Algorithms

The second article, Why So Many Clustering Algorithms, was chosen to delve into the complexities of clustering in order to highlight the nuances of the clustering problem. Estivill-Castro begins this article much like the previous one. Definitions are considered and notation is provided but this article focuses on more of the ambiguities of these areas. For each problem, the cluster thresholds, choosing to maximize, minimize, or calculate log-likelihood, cluster validity, and other parameters are scrutinized. The author points out that each of these parameters just mentioned are completely different focuses. So many design choices must be made for each problem that the number of algorithms stem from the vagueness that comes from these variant parameters [1]. This paper also points out that even in the same data set, the right clusters are ambiguous even from human to human and that the correct clusters are actually context dependant. So, to achieve validity, a validity index needs to be calculated for a problem using a baseline algorithm and then to compare effectiveness of one algorithm to another, all that needs to be considered is a comparison of the calculated validity indices [1]. However, even this presents a problem because the validity index would be biased on the induction principles used to calculate that index [1]. This paper did not answer any questions in as much as presented the complexities of the clustering problem. A more all encompassing viewpoint is examined in the next article.

4.3 Unification

Lastly, the article A Unified Framework for Model-based Clustering was selected to see if the nuances highlighted in the second article had been overcome or if the framework presented in this article was just unified for a particular problem set, as so many other models do. In this article, the author proposes a model based clustering that should work in a variety of contexts and be comparable if not better than the algorithms used to generate good clusters. They use different algorithms inside the model based on what the model is doing at the time: partitioning, clustering, or calculating stop conditions. The overarching method of this model is an annealing method. This method uses a temperature variable that is the probability that a new cluster should be started [2]. However, there are different ways to use the model based on the context of the data being examined [3], which requires that some knowledge of the data is known earlier. This is common to model-based algorithms in which the data is really being fitted to the model [1]. The only way to overcome the problem of no prior knowledge is to use a hierarchical model as well such that the knowledge is not needed. However, this increases the computational time drastically for large data sets [3]. This makes this framework too costly in those situations. The article goes on to show that the framework can be implemented differently based on the context of the problem set. Data is constructed and real life data is examined to show this flexibility in the proposed model. But how do these papers compare to each other and are there any differences in their claims?

5 Implications

These papers were very consistent in their examination of the clustering problem. A fundamental problem that they all agreed upon was that there is no real way to tell how many clusters an algorithm should end with. There are always stopping criteria but they are affected by training order (in some cases), a calculated threshold, or constant threshold [2]. They also agree that having a human start the algorithm off with knowledge about where to look for clusters, the likely size of clusters that exist, or some other knowledge about the data drastically improves the effectiveness of the algorithm. In the way of cluster validity, Zhong uses different calculations depending on the various problems [3], but this too is a design aspect of the unified model that was proposed which invariably puts more importance on the researcher using the model than the model itself. All-in-all, the unified framework did seem to have more extensions than the sum of its parts. This probably stemmed more from the combination of algorithms used for the specific purpose they were

designed for such as partitioning, clustering, et cetera. Also, the data that was to be classified in one of Zhongs experiments stated that all classifications were correctly identified and the plot of the data color coded to the data was shown [3]. However, the classes that were colored did not seem to be in the right areas. In one of the graphs, data points were even missing. This shows two aspects that have been supported throughout the papers: clusters are ambiguous (dependant on the person looking at them) and validity is more complicated than simple heuristics can articulate.

Each model and algorithm come from the same overarching idea: clusters are based on similarities; most often similarity is based on the distance the points are from each other. The closer the points, the more likely they are in the same cluster. The biggest problems come from ambiguous clusters (largely overlapping) and continuous data (evenly spaced with little to no variation) [2], [1]. Largely overlapping clusters could be the emergence of a new subspecies in a population and continuous data could be the data points that make up the visible spectrum. The continuousness of the spectrum makes it very difficult for clustering algorithms to distinguish infrared from ultraviolet.

6 Conclusions

The three articles considered in this paper give a good overview of the state of clustering in research, the complexities of it, and the direction that researchers would like to go. Anecdotally looking at different titles and reading abstracts of more recent articles would seem to suggest that algorithms are either being created or changed to fit data specific problems. This approach to clustering seems to be retardant of the actual goal of the creating a clustering algorithm. A human can look at data, depending on the person and the data's form, and classify the data relatively easily. Humans find this easier to do with graphics than with raw data or tables. This ease is caused by humans' predisposition to viewing shapes and patterns by nature. Cause and effect relationships known about features in the data help experts in the field change initial parameters in clustering algorithms leading to better results. So, maybe if computers had more of an understanding of cause/effect relationships or created plots and used a sensor to visually examine data, they would be able to generalize more. In essence, it may be that artificial intelligence and machine learning as a whole are not to the complexity necessary to generalize clusters from data with such tools as of yet.

References

- [1] Estivill-Castro, V., "Why So Many Clustering Algorithms: A Position Paper," *ACM SIGKDD Explorations Newsletter*, Vol. 4, Iss. 1, (June 2002), pp. 65-75.
- [2] Jain, A., Murty, M., Flynn, P., "Data Clustering: A Review," *ACM Computing Surveys (CSUR)*, Vol. 21, Iss. 3 (Sept. 1999), pp. 264-323.
- [3] Zhong, S., Ghosh, J., "A Unified Framework for Model-based Clustering," *The Journal of Machine Learning Research*, Vol. 4, (Dec. 2003), pp. 1001-1037.