

# Unsupervised Classification Procedures Applied to Satellite Cloud Data

DIANA F. GORDON

*Navy Center for Applied Research in Artificial Intelligence  
Naval Research Laboratory, Code 5510  
4555 Overlook Avenue, S.W.  
Washington, D.C. 20375-5337*

AND

PAUL M. TAG

RICHARD L. BANKERT  
*Naval Research Laboratory, Code 7541  
7 Grace Hopper Avenue  
Monterey, CA 93943-5502*

Machine learning algorithms can be subdivided into two types, supervised and unsupervised. Supervised learning is the more useful technique when the data samples have known outcomes that the user wants to predict. On the other hand, unsupervised learning is more appropriate when the user does not know the subdivisions into which the data samples, using relevant predictor features, should be divided. Prior categorical division may not be obvious because the problem may be a new one, for which the user has little experience. In such a case, an unsupervised learning procedure can provide insight into groupings that may make physical sense and facilitate future analysis. In this report, we explore the potential of two unsupervised learning programs, AutoClass and K-Means, when applied to a data set that was developed from satellite imagery of cloud regions that were expertly labeled into ten classes. Because cloud types hold meteorological significance, an automated classification from satellite imagery is of obvious use. We compare cloud classes produced by these systems with traditional cloud classes.

*Key words:* unsupervised learning, clustering, pattern recognition, cloud classification

## 1. Introduction

Human problem solving is generally an exercise in studying input conditions to predict an outcome based upon previous experience with similar situations. Using a computer program for developing rules based upon a series of these experiences is called “supervised” learning. Supervised learning is used for data sets with cases having known outcomes; this type of learning is the more common form because data is usually collected with some outcome in mind. Unsupervised learning, on the other hand, is not guided — the classes for which data are available are not known. Such might be the case for a new problem for which the user has little experience. In this report, we apply two unsupervised learning programs, AutoClass and K-Means, to a meteorological data set developed from satellite imagery.

Our data set was developed from satellite imagery of cloud regions that were expertly labeled into ten classes (e.g., cumulus, cumulonimbus, cirrus). It is composed of 1633 labeled cloud samples, where each sample is a 16 by 16 pixel region containing at least 75% cloud covering. Each sample is represented by 204 calculated features, or attributes. These features fall into three main categories: textural (e.g., homogeneity, contrast), spectral (e.g., maximum pixel value, mean pixel value), and physical (e.g., mean albedo, cloud top temperature). Supervised learning methods have produced theoretical classification accuracies of nearly 80% for this data set [Bankert, 1994] (recent work, unpublished, has pushed overall accuracy to 87%). The purpose of this research is to compare cloud classes produced from unsupervised learning procedures to traditional cloud classes. The differences may provide insight into alternate ways of viewing or interpreting clouds.

Our primary motivations for this research were to answer the following questions:

- Manual classification of these cloud images took *four* meteorologists approximately 25 to 50 hours *each* [Crosiar, 1993]. Can we automate this process using unsupervised classification to reduce the human effort?
- Can AI systems find new classifications with meteorological significance?
- In many areas of science, there is little or no expertise available; therefore, clustering is a potentially useful tool. How well can *existing* clustering techniques perform in an area in which there is a known standard (e.g., meteorological cloud classification)?

To avoid confusion of terminology, in the remainder of this paper we use the term *class* to refer to the classification provided by an expert meteorologist and *cluster* to refer to the classification derived by one of the clustering systems.

## 2. The algorithms

We briefly describe K-Means and AutoClass in the following subsections.

### 2.1. K-Means

The original ideas behind the K-Means algorithm were presented in MacQueen [1967]. This algorithm, which was originally developed for statistical pattern recognition, is quite simple. Assuming you want  $k$  clusters, the value  $k$  is provided as a

parameter. Given this parameter,  $k$  instances are selected randomly to form the cluster centroids. Every instance is then assigned to its nearest centroid, where the Euclidean distance measure is used. The  $k$  centroids are then recomputed. Each new centroid of a cluster is the vector centrally located among all instances of the cluster. This procedure is repeated until cluster membership no longer changes each iteration. The algorithm is summarized in Figure 1.

- Randomly select  $k$  instances as cluster centroids.  
 Until membership of instances in clusters stops changing:
1. Compute distances between every instance and every centroid.
  2. Assign each instance to the closest centroid (cluster).
  3. Recompute cluster centroids.

FIGURE 1. K-MEANS ALGORITHM.

The best and most complete description of the K-Means algorithm can be found in Duda and Hart [1973].

## 2.2. *AutoClass*

AutoClass was originally developed by Cheeseman et al. [1988] and continues to be widely used. It is a Bayesian approach to unsupervised learning based on finite mixture distributions. An important distinction between AutoClass and other clustering systems is that AutoClass searches for the ‘‘best’’ set of cluster descriptions rather than grouping the cases themselves. Clusters are described in terms of probability distribution functions and the locally maximal posterior probability parameters. The classifications are rated with an approximate posterior probability of the distribution function with respect to the data, obtained by marginalizing over all the parameters. The objective is to find the most probable classification (clustering).

Although AutoClass has the advantage that the number of clusters can be determined automatically, in our experiments (described below) we do not use this feature but instead fix the number of clusters at ten, the number of classes used by meteorologists in the original cloud classification study. AutoClass has been tested on several large, real databases and has discovered previously unsuspected classes [Cheeseman et al. 1988]. The algorithm is briefly summarized in Figure 2. For more complete descriptions of the AutoClass system, see Hanson et al. [1991a;b].

## 3. Empirical results

We would like to emphasize that in running the two systems we were *not* performing a systems comparison. Our primary interest was in determining how well existing clustering methods perform on this task. Therefore, we report on the ‘‘best’’ results from all experimental runs. For AutoClass, what we report is the best out of 76 runs, where ‘‘best’’ is defined as the hypothesis that maximizes the posterior probability. For K-Means, we report the best out of ten runs, where ‘‘best’’ is defined as the hypothesis that minimizes the distances between instances and their nearest cluster centroid. In all cases we fixed the number of clusters at ten.

Unfortunately, the current performance metrics for unsupervised learning were inadequate for our needs. We therefore developed an original performance metric to supplement the more traditional one of reporting the number of classes found [Cheeseman et al. 1988]. Our original measure is that of *cluster purity*. Before explaining

Form of a hypothesis: A distribution function  $T$  with parameters  $V$ .

BAYES'S THEOREM:

$$P(T | DI) = \frac{\int dV P(T | I) P(V | TI) P(D | VTI)}{P(D | I)}$$

USER INPUT:

- $D$  - Database of instances (cases)
- $T$  - Number of clusters, distribution type for values of each attribute

$I$  is implicit information not specifically represented.

Specification of  $T$  implicitly defines  $V$ ,  $P(T | I)$ , and  $P(V | TI)$ .

$P(T | I)$  and  $P(V | TI)$  are the minimum information priors.

SYSTEM OUTPUT (for each classification returned):

- $P(T | DI)$  - probability of distribution function  $T$
- $V$  - Maximum posterior probability parameter set, which with  $T$  specifies the classification description
- $P(D_i \in C_j | VTI)$  - Probabilistic assignment of cases to clusters

ALGORITHM:

Do for as long as you can afford:

1. Choose the number of classes, and thus  $T$ .
2. Begin with a consistent randomly chosen parameter set  $V$ .
3. Repeat until  $V$ ,  $P(D | H)$  reach a steady state:
  1. Compute  $P(D | H)$ , use Bayes's Theorem to approximate  $P(T | DI)$ .
  2. Compute class probabilities  $P(D_i \in C_j | VTI)$ .
  3. Reestimate  $V$  from the class probabilities.
4. Save  $T$ ,  $V$ ,  $P(T | DI)$ ,  $P(D_i \in C_j | VTI)$ .

Report those  $T$ ,  $V$ ,  $P(T | DI)$ ,  $P(D_i \in C_j | VTI)$  having the largest  $P(T | DI)$ .

FIGURE 2. AUTOCLASS.

cluster purity, let us examine a sample of the results from K-Means in Figure 3. Each row of the table corresponds to a cluster (only Cluster 5 through Cluster 7 are shown) and each column of the table corresponds to a class (only five out of the ten classes used by meteorologists are shown). In the table, "CI" is cirrus, "NS" is nimbostratus, "SC" is stratocumulus, "ST" is stratus, and "CU" is cumulus. Each entry in the table represents the percentage of the instances (cases) of that cloud class that appear in each cluster. Note that if we had shown all of the clusters, every column would sum to 100%. Also note that the total percentage of all classes in each cluster may be greater or less than 100%. We refer to the *primary*, *second*, and *third* class for each cluster as the class having the highest, second highest, and third highest percentage of cases, respectively. A class may be the primary, second, or third for more than one cluster.

Then the average cluster purity measure is the following:

$$\text{Average over all clusters of } \frac{\% \text{ of primary class in cluster}}{\text{total \% of all classes in cluster}} \times 100$$

For example, the average purity of Cluster 5 in Figure 3 is  $97/114 \times 100$ .

	CI	NS	SC	ST	CU
CLUSTER 5	16%	97%	1%	0%	0%
CLUSTER 6	54%	0%	1%	1%	7%
CLUSTER 7	15%	3%	0%	0%	0%

FIGURE 3. CLUSTERING RESULTS FROM K-MEANS.

Using our purity measure Figures 4 and 5 show the results with K-Means and AutoClass. Figure 4 groups by clusters. We can get a rough idea of the purity by the amount of white in the histogram; however, above each histogram we can see the precise average purity. K-Means had an average purity of 63.9% whereas AutoClass had an average purity of 47.8%.

Figure 5 groups by the ten cloud classes. From these histograms we can get an idea of what percentage of each cloud type appears as primary in each cluster. For K-Means, eight out of ten cloud classes are primary in a cluster. For AutoClass, six out of ten cloud classes are primary in a cluster. It is also interesting to observe that cirrocumulus does not appear as primary in any cluster in either K-Means or AutoClass. This omission may result from a combination of cirrocumulus having the second fewest data samples among the 1633 cloud samples and the fact that cirrocumulus is subtly different from cirrus and cirrostratus.

In general, both systems performed well, especially K-Means. We conclude that existing clustering techniques have the potential to automate important meteorological classifications; however, further investigation is needed.

#### 4. Future work

From the experiments performed with the K-Means and AutoClass systems on our cloud database, we see a good potential for existing clustering techniques to automate the process of classification of certain scientific data, thereby reducing the human effort required. K-Means appears especially promising, though we plan to continue exploring alternative parameter settings with AutoClass as well. Although neither system discovered new clusterings of meteorological significance, we would like to pursue this question further in the future. Furthermore, we would like to analyze the reasons for the differences in the results of the two systems, do more runs to collect statistics, and improve the algorithms (or, as in the case of AutoClass, improve our choice of parameter settings). Additionally, we plan to run other clustering systems, such as COBWEB [Fisher, 1987], on this database to see if we can get even better results. We would also like to augment our purely automated approach with methods for visualizing the data in order to better understand the characteristics of the database. Finally, we plan to try these techniques on other meteorological databases of interest.

#### Acknowledgements

The authors are grateful to Behzad Kamgar-Parsi for running (and explaining) his K-Means algorithm code, Will Taylor for running AutoClass for us, and John Stutz for patiently helping us to understand AutoClass.

## References

- BANKERT, R. L. 1994. Cloud classification of AVHRR imagery in maritime regions using a probabilistic neural network. *Journal of Applied Meteorology*. **33** : 909-918.
- CHEESEMAN, P., KELLY, J., SELF, M., STUTZ, J., TAYLOR, W., and FREEMAN, D. 1988. AutoClass: A Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, pp. 54-63.
- CROSIAR, C. L. 1993. An AVHRR cloud classification database typed by experts. Naval Research Laboratory Memorandum Report NRL/MR/7531-93-7207, 7 Grace Hopper Avenue Stop 2, Monterey, CA 93943-5502, 31 pp.
- DUDA, R. and HART, P. 1973. *Pattern Classification and Scene Analysis*. New York: John Wiley and Sons.
- FISHER, D. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*. **2/2** : 139-172.
- HANSON, R., STUTZ, J., and CHEESEMAN, P. 1991a. Bayesian classification theory. NASA Ames Research Center Technical Report FIA-90-12-7-01.
- HANSON, R., STUTZ, J., and CHEESEMAN, P. 1991b. Bayesian classification with correlation and inheritance. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pp. 692-698.
- MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, I, pp. 281-297.