Application of Machine Learning Clustering Algorithms to Thematic Map Design

Mark Polczynski – Marquette University
Michael Polczynski – Georgetown University

Key Words: Thematic maps, choropleth maps, data mining, machine learning, clustering algorithms.

Abstract
The application of data mining and machine learning technology and especially machine learning clustering algorithms can aid in the design of thematic maps portraying structural characteristics of geographical distributions. However, a review of the literature indicates that such applications are relatively rare. In order to stimulate greater use of clustering algorithms for thematic map design, an overview of the \textit{k means} and \textit{EM} clustering algorithms is given and two illustrative examples of the application of these algorithms to thematic map design are offered. The first example provides a choropleth map of the United States with states classified by two characteristics: population density and population growth. The second example identifies geographical clusters of sites associated with the 17\textsuperscript{th} century frontier between the Polish-Lithuanian Commonwealth and the Ottoman Empire. An overview of further applications of machine learning to characterization of this frontier region is summarized, and an appendix provides additional detail on the \textit{k means} clustering algorithm as applied in the first example provided here.

Introduction
Computer technology is having a major impact on both the design and production of maps. This is particularly true for thematic maps, which portray the structural characteristics of some particular geographical distribution not apparent in data presented in tabular form (Robinson 1975, 9-14). Coupled with the general availability of databases containing massive amounts of digital information, computer technology has made thematic maps an increasingly common element of media ranging from scholarly publications to daily news tabloids. Computer technology and the availability of massive databases have also enabled significant progress in the field of data mining and machine learning, which is also focused on finding and describing structural patterns in data (Witten, Frank and Hall 2011, 5-9). Data mining and machine learning are now widely applied in areas ranging from internet search engines to manufacturing quality improvement (Polczynski and Kochanski, 2010).

Given that both thematic maps and data mining and machine learning are focused on communicating the structure of data in a form readily comprehensible by humans, there exists a natural affinity between these two fields. The purpose of this brief technical note is to help stimulate the expansion of the application of data mining and machine learning technologies to the design of thematic maps.

This note contains the following sections: The research question frames the scope of this work, which is the application of machine learning clustering algorithms to the design of thematic maps illustrating structural characteristics of geographical distributions. A focused literature review traces selected highlights in the application of computers and especially machine learning to the design of thematic maps. A brief overview of two clustering methods is provided, and two illustrative examples of the application of clustering to map design are offered. Each example contains a brief description of the data used to design the map, the method of data analysis, and the result of applying clustering to the data presented as a thematic map. The conclusion describes how the examples answer the research question, and outlines a specific area for continued expansion of the application of data mining and machine learning to thematic map design. An appendix provides additional detail on one common clustering algorithm.

Research Question
The research question addressed here can be broken down into three levels:

1. Can data mining and machine learning technology be used as a general tool to aid in designing thematic maps?
2. Can machine learning clustering algorithms be used to discover structural characteristics of geographical distributions?
3. Does the \textit{Weka} data mining and machine learning workbench provide an appropriate toolset for cartographers preparing thematic maps?
Regarding the first question, one objective of thematic mapping is to portray the structural characteristics of some particular geographical distribution not apparent in data presented in tabular form. The objective of data mining is the extraction of implicit, previously unknown, and potentially useful information from data using machine learning algorithms designed to find and describe structural patterns in data. Thus, thematic maps and data mining share a common goal of converting databases into a format readily comprehensible by humans.

Given the long history of the application of computer technology to thematic map design and the general availability of robust data mining and machine learning tools, it would seem that these common goals would result in widespread application of machine learning technology to thematic map design, yet examples of this are rare compared to the overall volume of research done on thematic mapping. This leads to the highest level research question posed here: Can machine learning technology be used as a general tool to aid in designing thematic maps?

Clustering is one of the basic types of machine learning tools used for data mining. Clustering algorithms examine the attributes of a set of objects, and then separate the objects into different clusters with similar attributes. By analogy to thematic maps, the objects being clustered correspond to the particular geographical distribution being mapped, and object attributes correspond to the structural characteristics of the distribution. With specific reference to choropleth maps, the clusters formed correspond to the classes of the geographical distribution, and the process of assigning objects to clusters is termed classification\(^1\). This analogy leads to the next level of research question posed here: Can clustering algorithms be used to classify geographical distributions?

Affirmative answers to the preceding questions are irrelevant if data mining and machine learning tools suitable for application to thematic map design are not readily-available to map designers. Weka is a workbench of data mining and machine learning algorithms developed at the University of Waikato\(^2\). In spite (or possibly because) of the fact that Weka is free and open-source, the tools are robust and are widely-used by the academic community. User manuals, training materials, and tutorials for Weka are readily available via the web. Given that Weka is readily available to map designers leads to the last question posed here: Does the Weka data mining and machine learning workbench provide an appropriate toolset for cartographers preparing thematic maps?

**Literature Review**

The literature on the use of computers for map design and production is quite extensive. For example, since 1974 the annual AutoCarto Symposium on Automated Cartography\(^3\) has generated hundreds of publications in this area. Development of geographic information systems (GIS) has significantly added to this literature, with one review finding 319 relevant publications in the specific area of GIS-based multicriteria decision analysis (Malczewski 2006). A similar situation applies to data mining and machine learning technology, where thousands of articles and books have been published over the last decades. Given the abundance of literature in these areas, this review focuses exclusively on selected examples of work lying at the crossroads of thematic map design and data mining and machine learning. Key references to publications that can aid cartographers in applying data mining and machine learning algorithms and the Weka toolkit in particular are also included.

Although not typically associated with machine learning, natural breaks (Jenks) classification provides an early example of an approach to choropleth map design which relies on iterative calculations that can only be practically implemented using a computer (Robinson and others 1984, 365; Dent 1985, 205). Dating back to 1967, natural breaks classification (Jenks 1967; Coulson 1987) bears a strong resemblance to k means clustering (MacQueen 1967), a commonly used machine learning algorithm which also dates back to this time period. The brief description of k means clustering provided later in this note reveals the similarities between k means clustering and natural breaks classification.

One example of the application of technologies commonly associated with machine learning to thematic map design is ChoroWare, a software toolkit for choropleth map classification (Armstrong, Xiao, and Bennett 2003; Xiao and Armstrong 2006). The specific issue addressed by ChoroWare is that choropleth map design is typically constrained by multiple conflicting objectives, making selection of map class intervals a multiojective problem. Choroware uses genetic algorithms to generate a set of nondominated solutions to the multiojective choropleth class interval problem. Choroware is free open-source software available for download\(^4\), and runs under the Unix

---

\(^1\) As is commonly the case when fields of study with different origins intersect, a clash of terminology occurs. Machine learning clustering corresponds to choropleth map classification. The difficulty that arises is that in machine learning terminology, classification refers to a completely different type of data mining function.


\(^3\) AutoCarto Symposium on Automated Cartography: [http://www.cartogis.org/autocarto.php](http://www.cartogis.org/autocarto.php)

\(^4\) Available at [http://choroware.sourceforge.net/](http://choroware.sourceforge.net/)
operating system. Genetic algorithms comprise a set of techniques commonly used in machine learning for optimization and search problems (Coley 2003).

Recent work by Andrienko and Andrienko views computational methods associated with data mining and machine learning as a compliment to graphical methods, with the objective being to gain additional knowledge about data which cannot be easily acquired directly from viewing and manipulation of graphics (Andrienko and Andrienko 2006, 396-415). This work refers specifically to the application of machine learning clustering and classification algorithms to the design of choropleth maps.

As noted, the Weka workbench is used here to illustrate the application of data mining and machine learning technology to thematic map design. The researchers primarily responsible for developing Weka have provided a resource which covers both data mining and machine learning technology and the use of Weka (Witten, Frank and Hall 2011). A listing of over 300 academic publications that use Weka is available on-line5, as are numerous Weka tutorials6. Cios et. al. provide a useful general reference on data mining and machine learning (Cios and others 2007).

Method

Two different clustering algorithms are used for the examples provided here: the k means algorithm (Hartigan 1975, 84-107) and the EM (expectation maximization) algorithm (Dempster, Laird, and Rubin 1977). The k means algorithm is an iterative procedure which repeats until an optimum clustering of objects has been achieved. Listing 1 provides a general description of the k means algorithm. Additional detail on how k means clustering is performed can be found in the Appendix. The EM algorithm is more complex than k means and a detailed description of algorithm operation is beyond the scope of this work, but in essence the algorithm is similar to k means except that instead of assigning each object to a specific cluster, EM treats each cluster as a probability distributions and iteratively shifts objects among clusters to optimize the probability that each object belongs to its associated cluster (Han and Kamber 2006, 429-431). Thus, EM cluster boundaries are “fuzzy” vs. the “hard” boundaries created by k means clustering.

Listing 1: General description of the k means clustering algorithm (see Appendix for details).
1. Specify the desired number of clusters.
2. Guess the center of each cluster by choosing a random value for each object attribute.
3. For each object, calculate the distance from the object to the center of each cluster.
4. Assign each object to the nearest cluster.
5. Re-calculate the center of each cluster.
6. Go to Step 3 until no objects change cluster.

Multi-Characteristic Choropleth Map of the United States

For this first illustrative example of the application of machine learning clustering algorithms to thematic map design, clustering is used as an alternative to a natural breaks classification procedure for preparing choropleth maps. As with natural breaks classification, the objective of clustering is to classify spatial objects by placing class boundaries in the largest gaps between ranked objects with the intent of providing a high degree of similarity among all the objects within a particular class. Unlike typical implementations of natural breaks classification, machine learning clustering algorithms can form clusters using multiple object characteristics. This example uses clustering to produce a choropleth map for the 48 contiguous United States based on multiple demographic characteristics.

Data: For this example, two characteristics were used to perform clustering on the states: population density in 2008, and percent population change from 2000 to 2008, both obtained from the U.S. Census Bureau.

Method: The particular clustering algorithm used for this example is k means clustering, implemented as simple k means in the Weka workbench (Witten, Frank and Hall 2011, 139-141). For the illustrative example being followed here, five clusters were formed using the two attributes of population density and percent population change.

Results: Figure 1 shows a plot of population density vs. population change for the 48 states, and Figure 2 shows a choropleth map for the five clusters formed by the k means algorithm when clustering states by both population density and population change. Figure 1 uses different plot symbols to show the five clusters of states. The lightest shade in Figure 2 is used for states with high density and low growth (circles in lower right portion of Figure 1), and the darkest shade is used for states with low density and high growth (triangles in upper left portion of Figure 1).

5 Available at http://www.cs.waikato.ac.nz/ml/publications.html
6 See, for example: http://www.technologyforge.net/WekaTutorials/
Clusterings of Sites on the Polish-Lithuanian Commonwealth / Ottoman Empire Frontier

In this second example of the application of machine learning clustering algorithms to thematic map design, clustering is used as a means of helping to reveal structural characteristics of the 17th century Polish-Lithuanian Commonwealth – Ottoman Empire frontier. During this period, the Commonwealth/Empire border stretched 1,200 km from the Pontic Steppe of Eastern Ukraine to the Tatra Mountains in the Central Europe, making it the most extensive contiguous land-based border shared by a European and an Islamic power during the Early Modern period. This geographic setting played host to multiple coinciding frontiers, including: confessional (Christian Latin / Christian Orthodox / Muslim / Jewish), political (Commonwealth / Sultanate), lifeway (settled agriculturalism / pastoral nomadism), and environmental (deciduous forest / steppe). The 17th century was a period of continuous conflict and shifting of territorial ownership and allegiance, which caused the frontier between these two powers to be in a constant state of flux. Characterization of the structure of this frontier can be facilitated through the preparation of thematic maps.

Data: One way to characterize the geographical structure of the Commonwealth/Empire frontier is through mapping of frontier sites (cities, fortresses, etc.). The database being developed by one of the authors to characterize this frontier currently contains 728 sites. The database includes attributes such as site location (latitude and longitude), site type (e.g., fortified town, fortress, citadel, etc.), and construction type (e.g., earthworks, wood, stone). For the illustrative example provided here, clustering is applied to site latitude and longitude only to identify groupings of sites that are geographically near to each other relative to other sites along the frontier. For this work-in-progress, 119 sites have had exact latitude and longitude verified through visual inspection of the Google Earth map terrain in the vicinity of the site locations given by the primary sources.

Method: The EM clustering algorithm is used for this example. Unlike the k means algorithm which requires the number of clusters to be specified before clustering, the Weka implementation of EM can be configured to automatically identify the optimal number of naturally-occurring clusters in the dataset (Witten, Frank, and Hall 2011, pp. 287-288).

Results: Based on geographical location, the EM algorithm found three naturally-occurring clusters in the frontier site database. These optimal cluster assignments were then added to the frontier site database, and Figure 3 was created in Google Earth from the database using an application developed by the authors (Polczynski and Polczynski, 2013).

Conclusion

Two thematic maps were designed using the Weka workbench, a free, robust, readily-available toolkit of data mining and machine learning algorithms with extensive support documentation. The databases used to design the maps were in comma-separated values format (.csv), which is supported by all commonly-used spreadsheet applications. Returning to the research questions posed here, the results generated indicate that the Weka workbench provides a useful toolset for cartographers preparing thematic maps (research question 3). For the second map designed here, a clustering algorithm found three naturally-occurring classes of objects based on object location, thus supporting the assertion that clustering algorithms can be used to discover structural characteristics in geographical distributions (research question 2). Given that the examples utilized just two versions of one of several types of machine learning tools available in Weka, and that Weka represents just one of many data mining and machine learning systems currently available to map designers, this work provides evidence that machine learning technology can be used as a general tool to aid in designing thematic maps (research question 1).

Based on these results, a future research question suggests itself: Will application of approaches and algorithms commonly associated with data mining and machine learning become de rigueur for thematic map designers? An example of a clear indication of this possibility would be inclusion of clustering algorithms such as k means in widely-used mapping systems such as ArcGIS, which currently provides natural breaks, equal interval, defined interval, quantile, and standard deviation classifications methods for choropleth map design (Ormsby and others 2001, 134-142).

Regarding specific future work by these authors, additional sites and site attributes are being added to the 17th century Polish Commonwealth / Ottoman Empire frontier database, and various clustering algorithms are being applied to the database to reveal more detailed structural characteristics of the frontier. Given the resulting clusters of frontier sites, a second phase of the research will employ a second major data mining and machine learning

7 For example, Guillaume Le Vasseur, Sieur de Beauplan’s 1660 Description d’Ukraine manuscript and maps.
8 Available for download at: http://www.technologyforge.net/XLS2KML
9 See: http://www.arcgis.com
function termed *classification* to determine the specific site attributes that form the basis for naturally-occurring clusters of sites. For this work, *Weka* classifier algorithms (Witten, Frank, and Hall 2011, 191-215) will be used to generate decision trees and rule sets that will allow classification of newly-identified sites based on site attributes. A particular area of interest will be application of *Weka* to map time-based shifts\(^\text{10}\) in the structure of the frontier.

**Appendix**

The example of clustering of U.S. states by two demographic attributes can be used to provide additional insight on how the k means clustering algorithm can be implemented. After specifying the desired number of state clusters in Step 1 of Listing 1 (five clusters in this example), the purpose of Step 2 is to set an initial center point for each cluster. This is done by choosing five values at random between the minimum and maximum values of population density for all 48 states. These values are randomly paired with five random values between the minimum and maximum values of change in population of the states to create the initial two-attribute center points of the five clusters.

For Step 3, the distance from each state to the center of each cluster is calculated. There are a number of ways to do this, but the most common is to calculate the Euclidean distance between these points. For this two-attribute example, the Euclidean distance \(d_{ij}\) from state \(i\) to cluster center \(j\) is:

\[
d_{ij} = \sqrt{(pd_i - cpd_j)^2 + (cp_i - ccp_j)^2}
\]

where \(pd_i\) is the population density of state \(i\), \(cpd_j\) is the center point of population density for cluster \(j\), \(cp_i\) is the change in population of state \(i\), and \(ccp_j\) is the center point of change in population for cluster \(j\). (The values for \(cpd_j\) and \(ccp_j\) are calculated in Step 2). Given the distance \(d_{ij}\) between each state and each cluster center, Step 4 finds the cluster center closest to each state and assigns the state to that cluster.

Step 5 then uses all of the states assigned to each cluster to calculate a new center point for the cluster. This is done by calculating the mean value of the attributes of the states assigned to the cluster. For this example, this means calculating the mean of the population density values of all states assigned to cluster 1 and the mean of the change in population of states assigned to cluster 1 to form the new center for cluster 1, and then repeating this for the four remaining clusters.

At this point the algorithm loops back to Step 3. Since Step 5 moves the cluster centers, the distances from each state to each cluster center changes in Step 3. This, in turn, can cause states to be assigned to different clusters in Step 4, which then requires re-calculation of cluster centers in Step 5. The algorithm continues to loop through Steps 3-5 until no states change clusters in Step 4.

\(^{10}\) [http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka](http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka)
References


Figure 1: Population density vs. population change for 48 contiguous states.

Figure 2: Choropleth map for five classes of states clustered by population density and population change.
Figure 3: Geographic clustering of 17th century Polish-Lithuanian Commonwealth / Ottoman Empire sites.