Case Study: Clustering Towns

The Boston Globe is one of two major dailies serving Boston and the surrounding area of eastern Massachusetts and southern New Hampshire. The *Globe* is the leading circulated newspaper in Boston with daily circulation of over 467,000 in 2003 compared to 243,000 for the *Boston Herald*, the other major daily in town. On Sundays, the *Globe* has circulation of over 705,000. Despite this leading position, in 2003 the *Globe* did not want to stand still. As with many newspapers, it faced declining readership in its core Boston market and strong competition from local papers in the suburban markets where some of its readers have migrated.

In order to compete better with the suburban papers, the *Globe* introduced geographically distinct versions of the paper with specialized editorial content for each of 12 geographically defined zones. Two days a week, readers are treated to a few pages of local coverage for their area. The editorial zones were drawn up using data available to the *Globe*, common sense, and a map, but no formal statistical analysis. There were some constraints on the composition of the editorial zones:

- The zones had to be geographically contiguous so that the trucks carrying the localized editions from the central printing plant in Boston could take sensible routes.
- The zones had to be reasonably compact and contain sufficient population to justify specialized editorial content.
- The editorial zones had to be closely aligned with the geographic zones used to sell advertising.

Within these constraints, the *Globe* wished to design editorial zones that would group similar towns together. Sounds sensible, but which towns are similar? That is the question that *The Boston Globe* brought to us at Data Miners.

Creating Town Signatures

Before deciding which towns belonged together, there needed to be a way of describing the towns—a town signature with a column for every feature that might be useful for characterizing a town and comparing it with its neighbors. As it happened, Data Miners had worked on an earlier project to find towns with good prospects for future circulation growth that had already defined town signatures. Those signatures, which had been developed for a regression model to predict *Globe* home delivery penetration, turned out to be equally useful for undirected clustering. This is a fairly common occurrence; once a useful set of descriptive attributes has been collected it can be used for all sorts of things. In another example, a long-distance company developed customer

signatures based on call detail data in order to predict fraud and later found that the same variables were useful for distinguishing between business and residential users.

TIP **Although the time and effort it takes to create a good customer signature can seem daunting, the effort is repaid over time because the same attributes often turn out to be predictive for many different target variables. The oft quoted rule of thumb that 80 percent of the time spent on a data mining project goes into data preparation becomes less true when the data preparation effort can be amortized over several predictive modeling efforts.**

The Data

The town signatures were derived from several sources, with most of the variables coming from town-level U.S. Census data from 1990 and 2001. The census data provides counts of the number of residents by age, race, ethnic group, occupation, income, home value, average commute time, and many other interesting variables. In addition, the *Globe* had household-level data on its subscribers supplied by an outside data vendor as well as circulation figures for each town and subscriber-level information on discount plans, complaint calls, and type of subscription (daily, Sunday, or both).

There were four basic steps to creating the town signatures:

1. Aggregation
2. Normalization
3. Calculation of trends
4. Creation of derived variables

The first step in turning this data into a town signature was to aggregate everything to the town level. For example, the subscriber data was aggregated to produce the total number of subscribers and median subscriber household income for each town.

The next step was to transform counts into percentages. Most of the demographic information was in the form of counts. Even things like income, home value, and number of children are reported as counts of the number of people in predefined bins. Transforming all counts into percentages of the town population is an example of normalizing data across towns with widely varying populations. The fact that in 2001, there were 27,573 people with 4-year college degrees residing in Brookline, Massachusetts is not nearly as interesting as the fact that they represented 47.5 percent of that well-educated town, while the much larger number of people with similar degrees in Boston proper make up only 19.4 percent of the population there.

Each of the scores of variables in the census data was available for two different years 11 years apart. Historical data is interesting because it makes it possible to look at trends. Is a town gaining or losing population? School-age population? Hispanic population? Trends like these affect the feel and character of a town so they should be represented in the signature. For certain factors, such as total population, the absolute trend is interesting, so the ratio of the population count in 2001 to the count in 1990 was used. For other factors such as a town's mix of renters and home owners, the change in the proportion of home owners in the population is more interesting so the ratio of the 2001 home ownership percentage to the percentage in 1990 was used. In all cases, the resulting value is an index with the property that it is larger than 1 for anything that has increased over time and a little less than 1 for anything that has decreased over time.

Finally, to capture important attributes of a town that were not readily discernable from variables already in the signature, additional variables were derived from those already present. For example, both distance and direction from Boston seemed likely to be important in forming town clusters. These are calculated from the latitude and longitude of the gold-domed State House that Oliver Wendell Holmes once called "the hub of the solar system." (Today's Bostonians are not as modest as Justice Holmes; they now refer to the entire city as "the hub of the universe" or simply "the Hub." Headline writers commonly save three letters by using "hub" in place of "Boston" as in the apocryphal "Hub man killed in NYC terror attack.") The online postal service database provides a convenient source for the latitude and longitude for each town. Most towns have a single zip code; for those with more, the coordinates of the lowest numbered zip code were arbitrarily chosen. The distance from the town to Boston was easily calculated from the latitude and longitude using standard Euclidean distance. Despite rumors that have reached us that the Earth is round, we used simple plane geometry for these calculations:

```
distance = sqrt(( hub latitude - town latitude)^2 + (hub longitude - town
longitude)^2)
angle = arctan((hub latitude - town latitude)/(hub longitude - town
longitude))
```

These formulas are imprecise, since they assume that the earth is flat and that one degree of latitude has the same length as one degree of longitude. The area in question is not large enough for these flat Earth assumptions to make much difference. Also note that since these values will only be compared to one another there is no need to convert them into familiar units such as miles, kilometers, or degrees.

Creating Clusters

The first attempt to build clusters used signatures that describe the towns in terms of both demographics and geography. Clusters built this way could not be used directly to create editorial zones because of the geographic constraint that editorial zones must comprise contiguous towns. Since towns with similar demographics are not necessarily close to one another, clusters based on our signatures include towns all over the map, as shown in Figure 11.12. Weighting could be used to increase the importance of the geographic variables in cluster formation, but the result would be to cause the nongeographic variables to be ignored completely. Since the goal was to find similarities based at least partially on demographic data, the idea of *geographic* clusters was abandoned in favor of *demographic* ones. The demographic clusters could then be used as one factor in designing editorial zones, along with the geographic constraints.

Determining the Right Number of Clusters

Another problem with the idea of creating editorial zones directly through clustering is that there were business reasons for wanting about a dozen editorial zones, but no guarantee that a dozen good clusters would be found. This raises the general issue of how to determine the right number of clusters for a dataset. The data mining tool used for this clustering effort (MineSet, developed by SGI, and now available from Purple Insight) provides an interesting approach to this problem by combining K-means clustering with the divisive tree approach. First, decide on a lower bound K for the number of clusters. Build K clusters using the ordinary K-means algorithm. Using a fitness measure such as the variance or the mean distance from the cluster center according to whatever distance function is being used, determine which is the worst cluster and split it by forming two clusters from the previous single one. Repeat this process until some upper bound is reached. After each iteration, remember some measure of the overall fitness of the collection of clusters. The measure suggested earlier is the ratio of the mean distance of cluster members from the cluster center to the mean distance between clusters.

It is important to remember that the most important fitness measure for clusters is one that is hard to quantify—the usefulness of the clusters for the intended application. In the cluster tree shown in Figure 11.13, the next iteration of the cluster tree algorithm suggests splitting cluster 2. The resulting clusters have well-defined differences, but they did not behave differently according to any variables of interest to the Globe such as home delivery penetration or subscriber longevity. Figure 11.13 shows the final cluster tree and lists some statistics about each of the four clusters at the leaves.

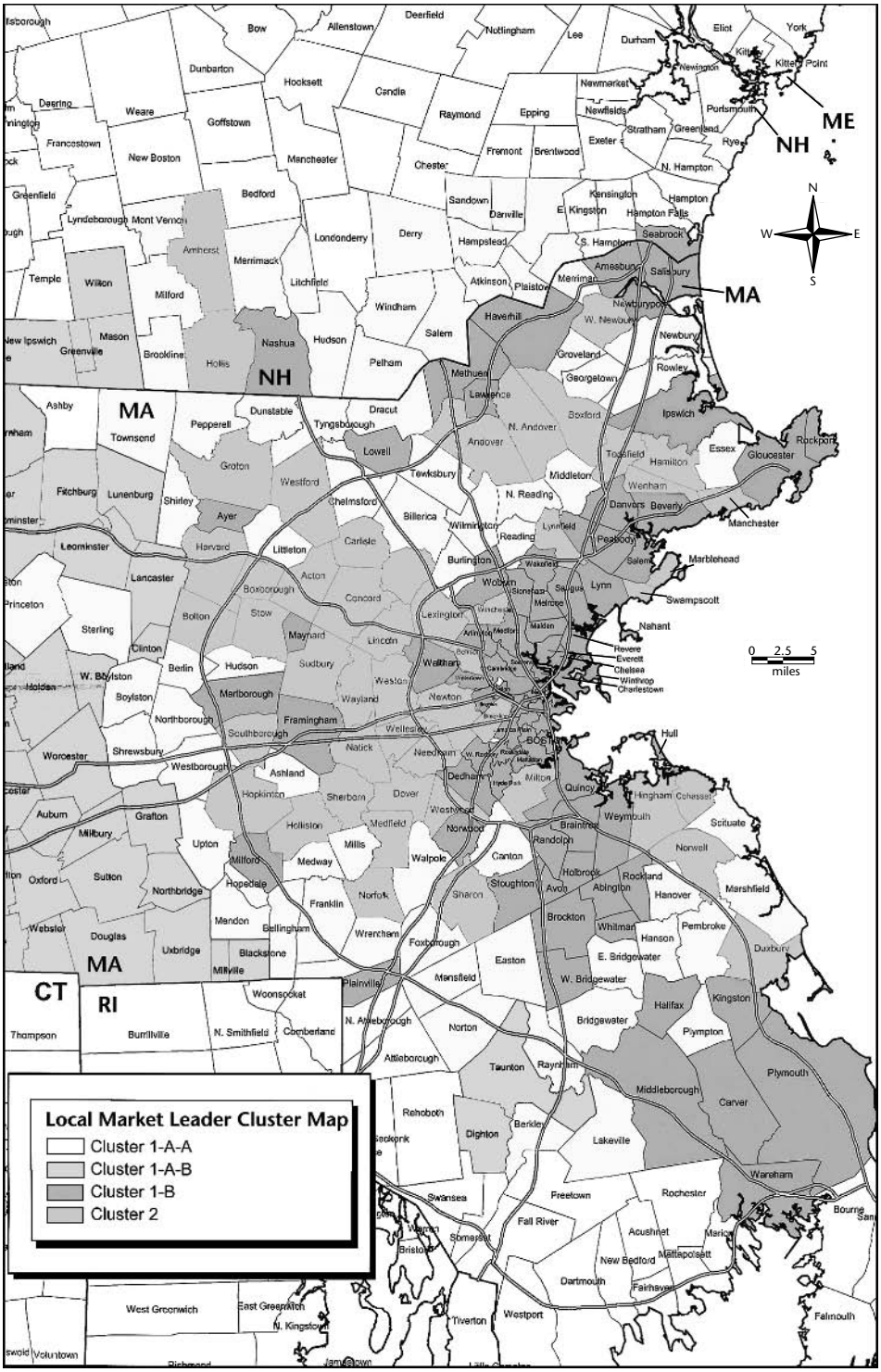

Figure 11.12 This map shows a demographic clustering of towns in eastern Massachusetts and southern New Hampshire.

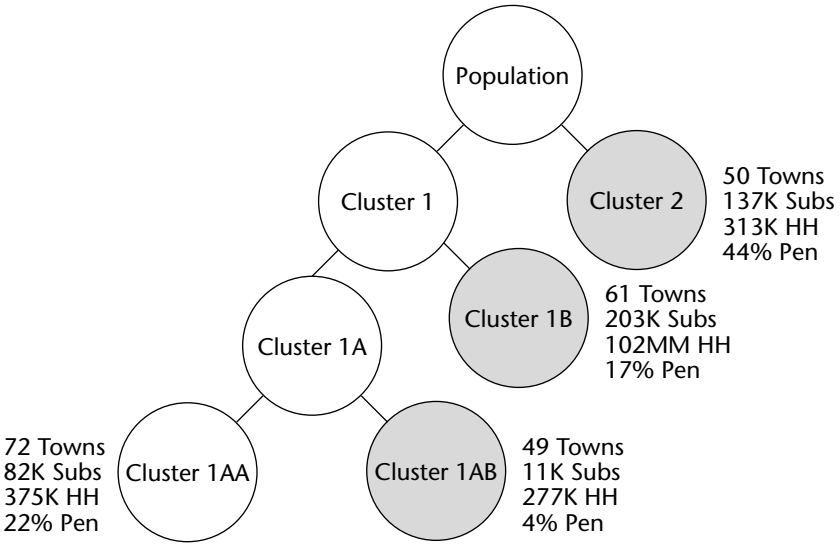

Figure 11.13 A cluster tree divides towns served by *The Boston Globe* into four distinct groups.

Cluster 2 contains 50 towns with 313,000 households, 137,000 of which subscribe to the daily or Sunday *Globe*. This level of home delivery penetration makes cluster 2 far and away the best cluster. The variables that best distinguish cluster 2 from the other clusters and from the population as a whole are home value and education. This cluster has the highest proportion of any cluster of home values in the top two bins; the highest proportion of people with 4-year college degrees, the highest mean years of education, and the lowest proportion of people in bluecollar jobs. The next best cluster from the point of view of home delivery penetration is Cluster 1AA, which is distinguished by its ordinariness. Its mean values for the most important variables, which in this case are home value and household income, are very close to the overall population means. Cluster 1B is characterized by some of the lowest household incomes, the oldest subscribers, and proximity to Boston. Cluster 1AB is the only cluster characterized primarily by geography. These are towns far from Boston. Not surprisingly, home delivery penetration is very low. Cluster 1AB has the lowest home values of any cluster, but incomes are average. We might infer that people in Cluster 1AB have chosen to live far from the city because they wish to own homes and real estate is less expensive on the outer fringes of suburbia; this hypothesis could be tested with market research.

Using Thematic Clusters to Adjust Zone Boundaries

The goal of the clustering project was to validate editorial zones that already existed. Each editorial zone consisted of a set of towns assigned one of the four clusters described above. The next step was to manually increase each zone's purity by swapping towns with adjacent zones. For example, Table 11.1 shows that all of the towns in the City zone are in Cluster 1B except Brookline, which is Cluster 2. In the neighboring West 1 zone, all the towns are in Cluster 2 except for Waltham and Watertown which are in Cluster 1B. Swapping Brookline into West 1 and Watertown and Waltham into City would make it possible for both editorial zones to be pure in the sense that all the towns in each zone would share the same cluster assignment. The new West 1 would be all Cluster 2, and the new City would be all Cluster 1B. As can be seen in the map in Figure 11.12, the new zones are still geographically contiguous.

Having editorial zones composed of similar towns makes it easier for the *Globe* to provide sharper editorial focus in its localized content, which should lead to higher circulation and better advertising sales.

Table 11.1 Towns in the *City* and *West 1* Editorial Zones

TOWN	EDITORIAL ZONE	CLUSTER ASSIGNMENT
Brookline	City	2
Boston	City	1B
Cambridge	City	1B
Somerville	City	1B
Needham	West 1	2
Newton	West 1	2
Wellesley	West 1	2
Waltham	West 1	1B
Weston	West 1	2
Watertown	West 1	1B

Lessons Learned

Automatic cluster detection is an undirected data mining technique that can be used to learn about the structure of complex databases. By breaking complex datasets into simpler clusters, automatic clustering can be used to improve the performance of more directed techniques. By choosing different distance measures, automatic clustering can be applied to almost any kind of data. It is as easy to find clusters in collections of news stories or insurance claims as in astronomical or financial data.

Clustering algorithms rely on a similarity metric of some kind to indicate whether two records are close or distant. Often, a geometric interpretation of distance is used, but there are other possibilities, some of which are more appropriate when the records to be clustered contain non-numeric data.

One of the most popular algorithms for automatic cluster detection is K-means. The K-means algorithm is an iterative approach to finding K clusters based on distance. The chapter also introduced several other clustering algorithms. Gaussian mixture models, are a variation on the K-means idea that allows for overlapping clusters. Divisive clustering builds a tree of clusters by successively dividing an initial large cluster. Agglomerative clustering starts with many small clusters and gradually combines them until there is only one cluster left. Divisive and agglomerative approaches allow the data miner to use external criteria to decide which level of the resulting cluster tree is most useful for a particular application.

This chapter introduced some technical measures for cluster fitness, but the most important measure for clustering is how useful the clusters turn out to be for furthering some business goal.