An order-specific clustering algorithm for the determination of representative demand curves

C.H. Marton, A. Elkamel *, T.A. Duever

Department of Chemical Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ont., Canada N2L 3G1

Received 24 January 2007; received in revised form 18 May 2007; accepted 14 June 2007
Available online 20 June 2007

Abstract

Data clustering consists of a group of procedures used to collect similar entries or data points within a set into clusters. No existing clustering technique considers entries sequentially in time. In some cases, it is desirable to generate clusters that represent a segment of a time-ordered data set. For these purposes, an order-specific clustering algorithm is proposed. The proposed algorithm employs representative load curves to describe the clusters it generates. The capabilities of the order-specific clustering algorithm are demonstrated on a case study using electricity demand data for the province of Ontario, Canada. Two different applications of the clustering algorithm on this data set are given to demonstrate the effect of error threshold values on the formation of clusters. An analysis of the error for each of these clustering applications is presented.

Keywords: Representative load curves; Similarity measure; k-Means clustering; Demand capacity planning

1. Introduction

Data clustering refers to a collection of techniques used to segment a large data set into a smaller number of groups, called clusters. The objective is to collect similar entries or data points in these clusters, such that intra-cluster similarity is maximized and inter-cluster similarity is minimized (Jain, Murty, & Flynn, 1999). Data clustering is employed in many different applications, such as gene expression (Cotta & Moscato, 2003; Garrigues et al., 2005), data mining (Tsai & Tsai, 2004), environmetrics (Dixon, 1994), and electricity demand characterization (Gasperic, Gerbec, & Gubina, 2002).

In many cases, the objective of data clustering is to reduce the amount of data to be analyzed by considering a representative point or curve of each cluster. From a much smaller set of these representative points or curves, it may be easier to identify the distinct features of each cluster, or the differences between clusters. Data clustering techniques can provide a tool to assist in identifying trends or characteristics of a data set.

There are a many different clustering techniques available, each with their own strengths and short-comings (Jain et al., 1999). It is important to note that data clustering is in most cases a means to an end, and not an end in itself. Selection of the appropriate clustering technique depends on the application, the data set, and the desired output. For a given application and data set, the best clustering technique depends on what type of trends the user wishes to capture.

In some situations, it may be desirable to segment a set of historical data into clusters, each describing a consecutive portion of time. Such a set of clusters could be used to characterize seasonal fluctuation. For example, if a chemical plant was to track product demand levels as a function of time for a year, clusters representing a specific block of time, such as a month or season, could be used to plan or schedule the required production in order to meet the demand (Sahinidis & Grossmann, 1991a, 1991b; Zentner, Elkamel, Pekny, & Reklaitis, 1998).

This article presents an order-specific clustering algorithm that can be used to identify seasonal trends in a set of time-ordered data, while preserving and accommodating daily fluctuations. This clustering algorithm is applied to a data set describing regional electricity demand as a function of time.

2. Data clustering overview

2.1. Similarity measure

The selection of an appropriate clustering algorithm depends on the method used to evaluate the similarity between different...
Fig. 1. Illustration of the integral of absolute error between two load duration curves.

data entries, points, or curves. There are many potential distance functions used to describe the dissimilarity between data points. The Minkowski metric provides a group of potential similarity measures of the form (between vectors $x_i$ and $x_j$):

$$d_p(x_i, x_j) = \left( \sum_{k=1}^{N} (x_{i,k} - x_{j,k})^p \right)^{1/p}$$  

where $N$ is the dimensionality of the data points and $p$ is the order of the Minkowski metric. As $d_p$ increases, $x_i$ and $x_j$ are less similar. Perhaps the most common form of this metric is the Euclidean distance function, where $p$ equals 2. Although commonly used, Euclidean distance can result in inadvertent scaling (Lavine, 2000). Common used distances are: Euclidean, Standardized Euclidean, Mahalanobis, Manhattan and Minkowski metric (Gasperic et al., 2002).

When the data to be clustered consists of a set of curves, some additional similarity measures may be used. A data set comprised of daily (e.g. demand) curves representing the measure of some value at each hour of the day could be considered as a set of data points in 24 dimensions. However, treating such a set as a group of curves may allow for a conceptually simpler analysis. One similarity measure that can be used to compare two curves is the integral of absolute error (IAE). An example of the physical interpretation of the IAE is shown in Fig. 1. The IAE describes the shaded region between the two daily curves, where the y-axis represents demand.

### 2.2. Clustering approach

Important characteristics of a clustering algorithm are the types of partitions made and the process by which the clusters are formed. When data points are placed in one and only one cluster, the partitions are termed hard. Clustering techniques which allow points to belong to more than one cluster at a time are said to have fuzzy partitions.

There are three main approaches to forming clusters. Divisive clustering techniques initially place all points in a single cluster, and then separate the set into smaller clusters based on the similarity measure. Agglomerative clustering techniques initially consider each point as a cluster, and then similar pairs of clusters are successively joined to produce larger clusters. Incremental clustering techniques assign the first point in a set to a cluster, then add the next point to the current cluster or start a new cluster based on a threshold value of the distance function, and repeat this process until all the data points have been considered.

### 2.3. Common clustering algorithms

Two basic clustering algorithms are $k$-means clustering (Jain et al., 1999) and combining hierarchical clustering (Gasperic et al., 2002). The $k$-means clustering technique uses hard partitions within the data set to produce $k$ clusters, where $k$ must be specified a priori. The $k$-means clustering algorithm proceeds as follows:

1. $k$ cluster centroid seeds are specified, either at random, or by selecting $k$ data points from the set.
2. Each data point in the set is assigned to the cluster to which the distance from the point to the center is the shortest. Commonly, the distance is calculated by the Euclidean distance function, but other distance functions may be used.
3. Once all points have been assigned to a cluster, the new centroid of each cluster is calculated as the average in each dimension of the members of the cluster.
4. Steps 2 and 3 are repeated until the cluster centroids move less than a specified threshold distance.

Combining hierarchical clustering is an agglomerative clustering technique that creates a hierarchical cluster tree to quantify the similarity between pairs of clusters. The combining hierarchical clustering algorithm proceeds as follows:

1. Each data point is initially assigned to its own cluster.
2. The distance between each pair of clusters is computed.
3. The pair of clusters with the smallest distance between the clusters is joined into a new cluster.
4. The new cluster centroid is calculated, and distances between the new cluster and the other remaining clusters are calculated.
5. Steps 3 and 4 are repeated until no additional agglomeration can be performed.

Once this process is complete, a figure which illustrates the distance between the various clusters is compiled. A sample of this cluster tree plot, called a dendrogram, is shown in Fig. 2. To make hard clusters using this technique, either the number of clusters desired or a threshold value for the distance must be specified. The cluster membership can be read from the dendrogram, where vertical lines refer to clusters, and horizontal lines indicate the joining of two clusters.

2.4. Representative load curves

In applications where daily fluctuation of a measured quantity is an important characteristic, there has been an interest in identifying representative load curves (RLC). Particularly, RLCs have been used to describe electricity demand (Balachandra & Chandru, 1999; Gasperic et al., 2002). A RLC shows a measured value at regular time intervals over the course of a day, and is representative of some larger group of days. Each of the two curves shown in Fig. 1 could be a RLC.

A review of the literature showed that RLCs have been generated using both statistical methods (multiple discriminant analysis) (Balachandra & Chandru, 1999) and clustering methods (Gasperic et al., 2002). However, the RLCs generated through a clustering method did not represent a period of time, but rather a group of businesses with similar needs. No clustering method that would consider entries sequentially in time has been proposed in the literature. In the next section, a clustering algorithm which identifies RLCs for groups that are continuous in time is proposed.

3. Order-specific clustering algorithm

The proposed data clustering algorithm is intended for use with a set of data with values at regularly spaced time intervals, specifically where daily fluctuations are expected. The objective of the clustering method is to generate RLCs that preserve order within a data set. The clusters generated by this proposed method will each represent a consecutive block of days within the set. The defining curve of each cluster will represent the average value at each time interval for a day. By grouping adjacent days, the clusters can be used to represent a known portion of time, and could be used to simplify the combinatorial complexity of decision-making models.

The incremental clustering approach is used in this clustering algorithm. The first day is assigned to the first cluster, and each time two consecutive days do not fit into the current cluster a new cluster is formed. The distance function used in this clustering algorithm is the IAE. The IAE is determined numerically using the Trapezoid rule and the multi-segment Simpson’s rule. The Trapezoid rule is required to integrate a single section if the curves have an even number of points.

Similar to the work by Balachandra and Chandru (1999), days that are determined not to match the larger cluster surrounding them are considered to be outliers. As described in the following cluster formation procedure, these outlier days are included in the cluster, although they do not contribute to the RLC. Certainly there are other variables which impact demand levels other than season or time of day, and this treatment of outliers is one way to accommodate the short-term disturbances these hidden variables may cause. In applications where a smaller number of clusters are desired for a relatively large data set, this treatment of outliers is essential to the success of the algorithm.

Similar to the collaborative fuzzy clustering algorithm used by Pedrycz (2002), our proposed algorithm uses two clustering passes to group the data. In this application, the cluster grouping procedure compares the localized, coarse-grained groups identified in the cluster formation procedure. The cluster grouping procedure can be thought of as looking for correlation in the data set over slightly larger time horizons than the cluster formation procedure.

3.1. Cluster formation procedure

Prior to beginning the cluster formation procedure, a threshold for the distance function must be specified. The following procedure assumes the IAE as the distance function. The threshold indicates the value for the distance function which is deemed an acceptable deviation for any cluster member from the cluster average. Selection of an appropriate threshold for similarity is subjective, and depends on any number of factors including the number of clusters sought, previous experience, and the goals of the study (Lavine, 2000). A flowchart representation of the cluster formation procedure is given in Fig. 3. In the flowchart, the error threshold is referred to as E1.

If it is desired to obtain a specific number of clusters from this algorithm, an iterative approach with respect to specifying the error threshold should be used. Specifically, select error
threshold values before using the clustering algorithm. Once the algorithm has been applied to the data, adjust the error threshold values based on the difference between the desired number of clusters and the actual number of clusters reported by the algorithm. Increasing the error threshold values will result in a smaller number of clusters.

The algorithm considers each daily curve from the data set, in sequence from the first day to the last. The procedure adds consecutive days to one cluster, provided they are deemed sufficiently similar, until consecutive days considered are not considered part of the cluster. When this occurs, the cluster is closed, and the algorithm moves forward starting a new cluster. This process is repeated until the algorithm reaches the end of the data set, at which point the last cluster is closed.

The initial clusters are formed according to the following procedure:

1. The first day is assigned to cluster 1, and the demand curve for that day is identified as the average curve for cluster 1.
2. The next day in the set is considered, and the IAE between the demand curve for the day being considered and the current cluster’s average curve is calculated.
3. The IAE is compared to the error threshold. There are three possible outcomes of this comparison:
   a. If the IAE is less than the error threshold, the day is added to the current cluster. The average curve for the cluster is recalculated to include the newly added day (i.e. a weighted average between the newly added demand curve and the old average curve is calculated such that all days in the cluster are equally weighted in the average). The current cluster, to which the day was added, remains the cluster considered in the next step.
   b. If the IAE is greater than the error threshold, the following day is compared with the current cluster. And if this second IAE is less than the error threshold, both days are added to the current cluster. The average curve for the cluster is updated (i.e. calculating a weighted average) only by the curve from the second day considered (where the IAE was less than the threshold). The first day considered, whose IAE was greater than the threshold, is marked as an outlier. The next day in the set is then considered for addition to the current cluster.
   c. If the IAE is greater than the error threshold, the following day is compared with the current cluster. And if this second IAE is greater than the error threshold, then the current cluster is closed, and the first of the 2 days considered is assigned to a new cluster (which becomes the “current” cluster). The second of the 2 days is then considered for addition to the new cluster in step 2.
4. Steps 2 and 3 are repeated until all the days in the data set have been considered.

3.2. Cluster grouping procedure

Prior to beginning the cluster grouping procedure, a second threshold for the distance function must be specified. The pur-
pose of this second procedure is to correct situations where a series of days that are either outliers or biased to one extreme of the cluster average create an artificial break within one cluster. A flowchart representation of this procedure is given in Fig. 4.

The cluster grouping procedure determines the IAE between the average curves of adjacent clusters, and compares the IAE to the second threshold value. If the IAE is less than the threshold, the clusters are grouped together (with a new average demand curve calculated from a weighted average) and this new cluster is compared to the next adjacent cluster. If the IAE is greater than the second error threshold value, no change is made to the clusters, and the next adjacent clusters are compared.

As with the first procedure, it may be beneficial in some cases to use an iterative process to determine an appropriate threshold value. The selection of the threshold value should be based on the observed trends in the clusters after the cluster formation procedure has partitioned the data set. In general, a larger threshold value will result in more clusters being joined, and a smaller threshold value will result in little or no change to the clusters formed in the previous step. In our initial application of this algorithm, we maintained the error threshold used in the cluster grouping procedure on the same order (or less) as the threshold value used in the cluster formation procedure. However, the impact of this rule of thumb was not specifically investigated.

The RLC for each cluster is the average demand curve once the cluster membership is finalized. Each value that makes up the RLC is an average demand level at that time point for all of the (non-outlier) days in the cluster.

4. Case study: Ontario electricity demand

As a demonstration, the proposed clustering algorithm was applied to electricity demand data from the province of Ontario from 1 November 2003 to 30 October 2004. This data was available from the Independent Electricity System Operator (IESO) of Ontario’s website (Hourly Demands, 2005). The data set gives the electricity demand for the province of Ontario for each hour of the 365 days, in megawatts (MW). The clustering tool was programmed in Visual Basic using Microsoft Excel.

The number of clusters formed using the order-specific clustering algorithm depends on the threshold values applied to the distance function (as with any clustering algorithm where the
Table 1
Threshold value influence on number of clusters

<table>
<thead>
<tr>
<th>Threshold Formation (MWh)</th>
<th>Threshold Grouping (MWh)</th>
<th>Number of Significant Clusters</th>
<th>Percentage of Days Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>30,000</td>
<td>40,000</td>
<td>11</td>
<td>71.2</td>
</tr>
<tr>
<td>40,000</td>
<td>50,000</td>
<td>9</td>
<td>92.9</td>
</tr>
<tr>
<td>50,000</td>
<td>60,000</td>
<td>7</td>
<td>96.2</td>
</tr>
<tr>
<td>60,000</td>
<td>70,000</td>
<td>6</td>
<td>99</td>
</tr>
</tbody>
</table>

number of clusters is not specified). For the case of Ontario electricity demand, several different threshold levels were applied. The different threshold values and the number of significant clusters formed are listed in Table 1. The term significant cluster was applied to clusters containing 10 or more days. The fraction of the year that is included in these significant clusters is also shown in Table 1. The clusters represented by the two shaded rows will be examined in detail below.

4.1. Clustering option (a): nine significant clusters

The error threshold values used to generate nine significant clusters, as shown in Table 1, were 30,000 MWh for the cluster formation procedure and 50,000 MWh for the cluster grouping procedure. As shown in Table 1, 92.9% of the days in the year were included in the nine significant clusters. The remainder of the data is neglected, as the curves were not similar to either of the two adjacent clusters (i.e. IAE not within the error threshold). The cluster membership is listed in Table 2. The RLCs for the nine significant clusters are shown in Fig. 5. Explanation of cluster RLC trends based on seasonal electricity usage is beyond the scope of this work. There seems to be a significant difference between the RLCs of adjacent clusters.

There are some features of the cluster membership and RLCs that suggest appropriate groups have been formed. Both the timing and RLC of cluster 2 seem to correspond with the winter holidays during which many businesses shut down or operate at a reduced capacity. The RLCs for clusters 7 and 8, which represent most of the summer, both exhibit a more sustained peak throughout the day, as opposed to the traditional curve with peaks in the morning and evening. This may represent the high demand for electricity used in air-conditioning.

The relationship between a RLC and the curves that make up a cluster is shown in Fig. 6. The figure shows each curve in cluster 2, as well as the RLC. The RLC appears to be a good approximation of these curves, particularly because the RLC and curves seem to exhibit a similar daily fluctuation.

In order to assess the validity of the clusters formed using this algorithm, an error histogram was prepared. The errors included were calculated according to the equation:

\[
\text{error}_{i,j,k} = \frac{D_{j,k} - \text{RLC}_{i,j}}{\bar{D}} \times 100
\]

where \(\text{error}_{i,j,k}\) is the scaled error for hour \(j\) of day \(k\), which is in cluster \(i\), \(D_{j,k}\) the electricity demand for hour \(j\) of day \(k\), \(\text{RLC}_{i,j}\) the value of the RLC for cluster \(i\) at hour \(j\) and \(\bar{D}\) is the average hourly demand for the entire data set.

Fig. 5. RLCs for Ontario energy demand formed under clustering option (a) with a cluster threshold of 30,000 MWh and a grouping threshold of 50,000 MWh.
This error represents the percentage of the average hourly demand for the year by which each day is different from the RLC by which it is represented. A histogram showing the distribution of the errors is shown as Fig. 7. From this histogram it appears that the errors are normally distributed. The mean error is $-0.08\%$, and the standard deviation is 6.53%. The mean and standard deviation of the error are similar to those obtained when generating nine RLCs from a year of electricity demand using multiple discriminant analysis (Balachandra & Chandru, 1999).

4.2. Clustering option (b): six significant clusters

The error threshold values used to generate six significant clusters, as shown in Table 1, were 60,000 MWh for the cluster formation procedure and 34,000 MWh for the cluster grouping procedure. As shown in Table 1, 99% of the days in the year were included in the six significant clusters. The cluster membership is listed in Table 3. The RLCs for the six clusters are shown in Fig. 8. As noted with the clusters formed in clustering option (a), there seems to be a significant difference between the RLCs of adjacent clusters.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Start date</th>
<th>End date</th>
<th>Size (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>1 November 2003</td>
<td>24 December 2003</td>
<td>54</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>25 December 2003</td>
<td>4 January 2004</td>
<td>11</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>5 January 2004</td>
<td>27 February 2004</td>
<td>54</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>28 February 2004</td>
<td>8 April 2004</td>
<td>41</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>9 April 2004</td>
<td>7 June 2004</td>
<td>60</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>8 June 2004</td>
<td>30 October 2004</td>
<td>140</td>
</tr>
</tbody>
</table>
A histogram showing the distribution of the errors is shown as Fig. 9. As with the histogram for clustering option (a), it appears that the errors are normally distributed. The mean error is 0.04%, and the standard deviation is 7.06%.

5. Conclusion

A data clustering algorithm that segments a time-ordered set of data into clusters representing a continuous span of time was proposed. The RLCs generated by the algorithm may be used to reduce the combinatorial complexity of decision-making models. The distance function used is the integral of absolute error between two curves. The proposed incremental clustering method follows a two-step process. The first step, called the cluster formation procedure, assigns the curves to different clusters. The second step, called the cluster grouping procedure, combines adjacent clusters that are determined to be sufficiently similar to one another. This clustering algorithm was applied to Ontario electricity demand data as a case study.

For the case of grouping a year of Ontario electricity demand curves into a small number of clusters, the order-specific clustering algorithm served as an adequate tool. Two different sets of clusters were presented; each generated using different error threshold values. A set of nine significant clusters was generated using error threshold values of 30,000 and 50,000 MWh for the cluster formation and cluster grouping procedures, respectively. A set of six significant clusters was generated using error threshold values of 60,000 and 34,000 MWh for the cluster formation and cluster grouping procedures, respectively. In both cases the error distribution appeared to be normally distributed. The set of nine clusters had a mean error of −0.08% and a standard error deviation of 6.53%. The set of six clusters had a mean error of 0.04% and a standard error deviation of 7.07%.

There are several different techniques that may be used to validate the clusters formed using a specific clustering algorithm (Chen et al., 2002; He, Tan, & Sung, 2002; Hibbs, Dirksen, Li, & Troyanskaya, 2005; Yeung, Haynor, & Ruzzo, 2001). However, most of these methods are meant to help select the best clustering algorithm for a specific application. A large number of clustering techniques have been developed because there are a wide range of applications for which different variations of clustering tools are most suitable. In most cases, clustering is more art than science, in that the goal is to identify groups that are deemed similar enough to suit some purpose. The order-specific clustering algorithm presented here can be used to generate RLCs that describe a continuous span of time.

Acknowledgement

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References


