White paper – technical report

The SPSS TwoStep Cluster Component

A scalable component enabling more efficient customer segmentation

SPSS®

Introduction

The SPSS TwoStep Clustering Component is a scalable cluster analysis algorithm designed to handle very large datasets. Capable of handling both continuous and categorical variables or attributes, it requires only one data pass in the procedure. In the first step of the procedure, you pre-cluster the records into many small sub-clusters. Then, cluster the sub-clusters from the pre-cluster step into the desired number of clusters. If the desired number of clusters is unknown, the SPSS TwoStep Cluster Component will find the proper number of clusters automatically.

The results gathered from running a simulation are consistently accurate and scalable in performance. The simulation also shows that the automatic procedure of finding the number of clusters works remarkably well and fast.

By clustering, you can group data so that records within a group are similar. For example, retail and consumer product companies regularly apply clustering techniques to data that describes their customers' buying habits, gender, age, income level, etc. These companies tailor their marketing and product development strategy to each consumer group to increase sales and build brand loyalty.

History of clustering methods

Traditional clustering methods fall into two broad categories: relocation and hierarchical. Relocation clustering methods — such as k-means and EM (expectation-maximization) — move records iteratively from one cluster to another, starting from an initial partition. You need to specify the number of clusters in advance, and it does not change during the iteration. Hierarchical clustering methods proceed by stages producing a sequence of partitions in which each one nests into the next partition in the sequence. Hierarchical clustering can be either agglomerative or divisive. Agglomerative clustering starts with singleton clusters (clusters that contain only one record) and proceeds by successively merging the two "closest" clusters at each stage. In contrast, divisive clustering starts with one single cluster that contains all records and proceeds by successively separating the cluster into smaller ones. Unlike relocation methods, you don't need initial values for hierarchical clustering.

All the aforementioned cluster methods (except the EM method) need a distance measure. Different distance measures may lead to different cluster results. Some distance measures accept only continuous variables like Euclidean distance, and some only categorical variables, such as the simple matching dissimilarity measure used in the k-modes method by Huang (1998). For mixed-type variables, various distance measures exist based on the weighted sum of continuous variables distances and categorical variables distances (Huang 1998, Kaufman and Rousseeuw 1990). You can choose the weight arbitrarily, but improper weight may bias the treatment of different variable types. Banfield and Raftery (1993) introduced a model-based distance measure for data with continuous attributes. They derived this measure from a Gaussian mixture model, equivalent to the decrease in log-likelihood resulting from merging two clusters. Meila and Heckerman (1998) applied this probabilistic concept and derived another distance measure for data with categorical attributes only. The SPSS TwoStep Cluster Component extends this model-based distance measure to situations that include both continuous and categorical variables.

Traditional clustering methods are effective and accurate on small datasets, but usually don't scale up to the very large datasets. These traditional methods will cluster large datasets effectively if these datasets are first reduced into smaller datasets. This is the basic concept of two-stage clustering methods like BIRCH (Zhang et al. 1996). In the first stage, you apply a quick sequential cluster method to the large dataset to compress the dense regions and form sub-clusters. In the second stage, apply a cluster method on the sub-clusters to find the desired number of clusters. The records in one sub-cluster should end up in one of the final clusters so the pre-cluster step will not affect the accuracy of the final clustering. In general, inaccuracy from the pre-cluster step decreases as the number of sub-clusters from the pre-cluster step increases. However, too many sub-clusters will slow down the second stage clustering. Choose the number of sub-clusters carefully so that the number is large enough to produce accurate results and small enough to not inhibit performance in the later clustering procedure.

None of the cluster methods directly address the issue of determining the number of clusters because the means of determining the number of clusters is difficult and it is considered a separate issue. You might apply various strategies to determine the number of clusters. You want to cluster the data into a series of numbers of clusters (such as two clusters, three clusters, etc.) and calculate certain criterion statistics for each of them. The one with the best statistic is the "winner." Fraley and Raftery (1998) proposed using the Bayesian information criterion (BIC) as the criterion statistic for the EM clustering method. Banfield and Raftery (1993) suggested using the approximate weight of evidence (AWE) as the criterion statistic for their model-based hierarchical clustering.

Improve your process with the SPSS TwoStep Cluster Component

With over 30 years of experience in statistical software, SPSS understands the advantages and disadvantages of other statistical methods and applied that knowledge to produce a new method. The SPSS TwoStep Cluster Component:

- Handles both continuous and categorical variables by extending the model-based distance measure used by Banfield and Raftery (1993) to situations with both continuous and categorical variables
- Utilizes a two-step clustering approach similar to BIRCH (Zhang et al. 1996)
- Provides the capability to automatically find the optimal number of clusters

Step 1: pre-cluster your data

The pre-cluster step uses a sequential clustering approach (Theodoridis and Koutroumbas 1999). It scans the records one by one and decides if the current record should merge with the previously formed clusters or start a new cluster based on the distance criterion. You implement the procedure by constructing a modified cluster feature (CF) tree (Zhang et al. 1996). The CF-tree consists of levels of nodes, and each node contains a number of entries. A leaf entry (an entry in the leaf node) represents a sub-cluster you want. The non-leaf nodes and their entries guide a new record into a correct leaf node quickly. For example, the SPSS default uses a CF-tree with a maximum of three levels of nodes and a maximum of eight entries per node. This combination may result in a maximum of 512 leaf entries, hence 512 sub-clusters.

A CF with the entry's number of records, the mean and variance of each continuous variable, plus the counts for each category of each categorical variable characterize each entry. Each successive record, starting from the root node, is recursively guided by the closest entry in the node to find the closest child node, then descends along the CF-tree. Upon reaching a leaf node, it finds the closest leaf entry in the leaf node. If the record is within a threshold distance of the closest leaf entry, the leaf entry absorbs it and updates the CF. Otherwise it starts its own leaf entry in the leaf node. If the leaf node has no space to create a new leaf entry, the leaf node splits in two. The entries in the original leaf node divide into two groups using the farthest pair as seeds, redistributing the remaining entries based on the closest criteria. If a CF-tree grows beyond the maximum number of levels, the CF-tree rebuilds the existing CF-tree by increasing the threshold distance criterion. The rebuilt CF-tree is smaller, so it has space for new input records. This process continues through a complete data pass. For more on CF-tree construction, see BIRCH by Zhang et al. (1996).

The CF used in this pre-cluster is different from the one used in BIRCH, which only handles continuous variables. BIRCH's CF comprises the entry's number of records, mean and variance of each continuous variable. The SPSS CF extends BIRCH's CF by including the counts for each category of each categorical variable. The entry's CF collectively represents all records falling in the same entry. When you add a new record to an entry, you can compute the new CF from the old CF without knowing the individual records in the entry. These properties make it possible to maintain only the entry CFs, rather than the sets of individual records. Therefore, the CF-tree is much smaller and more likely to be stored in the main memory.

Note: the CF-tree may depend on the input order of records. To minimize the order effect, use random order.

Step 2: group your data into sub-clusters

The cluster step takes sub-clusters resulting from the first step as input and then groups them into the desired number of clusters. Since the number of sub-clusters is much less than the number of original records, you can use traditional clustering methods effectively. SPSS uses the agglomerative hierarchical clustering method primarily because it works well with the auto-cluster procedure (see the auto-cluster section, which immediately follows). The component structure allows you to deploy future methods easily as they become available.

Apply the auto-cluster to determine the number of clusters

How many clusters are there? The answer depends on your dataset. Hierarchical clustering characteristically produces a sequence of partitions at one run: 1, 2, 3, … clusters. The k-means and EM method would need to run multiple times (one for each specified number of clusters) in order to generate the sequence. To determine the number of clusters automatically, SPSS developed a two-step procedure that works well with the hierarchical clustering method. The first step, calculates BIC for each number of clusters within a specified range and uses it to find the initial estimate for the number of clusters. The second step refines the initial estimate by finding the greatest change in distance between the two closest clusters in each hierarchical clustering stage.

Handle your data with a new distance measure

You need a distance measure in both the pre-cluster and cluster steps. In order to handle both continuous and categorical variables, define the distance between two clusters as the corresponding decrease in log-likelihood by combining them into one cluster. In calculating log-likelihood, assume normal distributions for continuous variables and multinomial distributions for categorical variables. Plus, assume that the variables are independent of each other, as well as the records.

Numerical and simulation studies

SPSS implemented the TwoStep Cluster Component in both Java and C++ language. It tested the performance of the TwoStep Cluster Component on simulated datasets. Results are presented in the table below. The total time used for each dataset shown here comes from a Java implementation using text input files run on a Pentium 800Mhz, 256MB RAM computer.

Datasets	Number of records (x1000)	Number of variables		True number of clusters	Number of sub-clusters by pre-cluster	No. of clusters found by auto-cluster	Percentage of wrongly clustered	Total time used (in seconds)
		Con	Cat*					
Data 1	200	2	0	5	199	5	0.09%	21
Data 2	400	2	0	5	269	5	0.03%	41
Data 3	500	2	0	5	207	5	0.35%	47
Data 4	1,000	2	0	5	297	5	1.2%	93
Data 5	2,000	2	0	5	243	5	0.07%	193
Data 6	2,500	2	0	5	187	5	0.11%	229
Data 7	8.4	640	0	7	71	7	0%	177
Data 8	1,000	0	5	4	243	4	0.03%	572
Data 9	1,000	5	5	8	232	8	0%	1,070
Data 10	1,000	25	25	10	264	10	0%	4,970

*All the categorical variables are of 12 categories.

Performance of the SPSS TwoStep Cluster Component

The two-step auto-cluster procedure developed by SPSS is different from any existing method. Simulation studies show that either BIC (or other criterion like AWE, AIC) or distance changes alone do not automatically find the number of clusters in many situations. Combining both BIC and distance change — as in our two-step auto-cluster procedure — works much better than using any one alone. For hundreds of simulated datasets that can be viewed graphically (two-dimensional or three-dimensional datasets), the auto-cluster procedure finds the number of clusters correctly.

Inputs and outputs of the SPSS TwoStep Cluster Component

Inputs

- Pre-cluster parameters
  - The input variables and their type (continuous or categorical): user must supply
  - The scale parameters for continuous variables, one for each continuous variable: User controlled. The default is the standard deviation of each continuous variable, if they are available. Otherwise the default is one.
  - Maximum number of levels of CF-tree: user controlled. The default is three.
  - Maximum number of entries per node: user controlled. The default is eight. Combining defaults of the maximum number of levels and the maximum number of entries per node may result in up to 512 sub-clusters. This quantity should be large enough to produce accurate results and small enough not to slow the later clustering procedure. Tested on thousands of datasets, these combined defaults are accurate and timely. Change these defaults only if you know different values would work better.
  - Size limit of main memory: user controlled. The default is set at 32MB, so that the whole CF-tree is more likely to fit in the main memory under reasonable conditions. If the CF-tree cannot fit in the main memory, it will spill onto the hard disk.
  - Size limit of storage space for outlier handling and delay-split: user controlled. The default is five percent of main memory size. (This default needs to be tested and modified).
  - Threshold distance: start with zero, then change to the smallest distance among CF records each time a CF-tree needs to be rebuilt. User cannot control.
  - Distance definition: 1 = distance defined by the log-likelihood decrease (this default is suitable for both continuous and categorical variables); 2 = average pair-wise Euclidean distance between two clusters (suitable only in combination with continuous variables). You can choose between these distances.

- Cluster parameters
  - Cluster method: currently only the hierarchical method is available. Future releases may integrate different clustering methods such as EM.
  - Number of clusters: user controlled. The default is auto-cluster.

- Auto-cluster parameters:
  - Range for number of clusters [Jmin, Jmax]: user controlled. The default is [1, 25].

Outputs

- Save any CF-tree to disk, if you want to update the model later

- Everything necessary to make membership assignment for new records, including:
  - Summary statistics for each final cluster
  - Clustering method and distance definition used in the clustering step
  - Scale parameters for each continuous variable
  - Criterion index: you can choose among none or BIC. The default is none. These indexes are useful when you want to find the proper number of clusters.

The SPSS advantage

Today's information age challenges developers to quickly create applications that transform data overload into critical information workers need to make smarter, faster decisions. Using SPSS' analytical technologies you can create applications in a fraction of the usual time and cost required — and with greatly reduced risk. Using SPSS' components ensures that your organization utilizes the most accurate, efficient and scalable analytical routines available in the marketplace. A knowledgeable support team of analytical and technical consultants is also available from SPSS to help you accelerate your development process even further — because every day that your application isn't in your employees or customers' hands, you lose potential utility, revenue and competitive advantage.

References

Banfield J. D. and A. E. Raftery. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49. p. 803–821.

Fraley C. and A.E. Raftery. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 4. p. 578–588.

Fraley, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20. p. 270–281.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2. p. 283–304.

Kaufman, L. and P.J. Rousseeuw. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley, New York.

Melia, M. and D. Heckerman. (1998). An experimental comparison of several clustering and initialization methods. *Microsoft Research Technical Report* MSR-TR-98-06.

Theodoridis, S. and K. Koutroumbas. (1999). *Pattern recognition*. Academic Press, New York.

Zhang, T., R. Ramakrishnon and M. Livny. (1996). BIRCH: An efficient data clustering method for very large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data.* p. 103–114, Montreal, Canada.

Data mining makes the difference

SPSS Inc. enables organizations to develop more profitable customer relationships by providing analytical solutions that discover what customers want and predict what they will do. The company delivers analytical solutions at the intersection of customer relationship management and business intelligence. SPSS analytical solutions integrate and analyze market, customer and operational data and deliver results in key vertical markets worldwide including: telecommunications, health care, banking, finance, insurance, manufacturing, retail, consumer packaged goods, market research and the public sector. For more information, visit **www.spss.com.**

Contacting SPSS

To place an order or to get more information, call your nearest SPSS office or visit our World Wide Web site at **www.spss.com**.

SPSS Inc.	+1.312.651.3000
Toll-free	+1.800.543.2185
SPSS Argentina	+5411.4814.5030
SPSS Asia Pacific	+65.245.9110
SPSS Australasia	+61.2.9954.5660
Toll-free	+1.800.024.836
SPSS Belgium	+32.16.317070
SPSS Benelux	+31.183.651.777
SPSS Brasil	+55.11.5505.3644
SPSS Czech Republic	+420.2.24813839
SPSS Danmark	+45.45.46.02.00
SPSS East Africa	+254.2.577.262
SPSS Federal Systems (U.S.)	+1.703.527.6777
Toll-free	+1.800.860.5762
SPSS Finland	+358.9.4355.920
SPSS France	+01.55.35.27.00
SPSS Germany	+49.89.4890740
SPSS Hellas	+30.1.72.51.925
SPSS Hispanoportuguesa	+34.91.447.37.00
SPSS Hong Kong	+852.2.811.9662
SPSS Ireland	+353.1.415.0234

SPSS Israel	+972.3.6506022
SPSS Italia	+800.437300
SPSS Japan	+81.3.5466.5511
SPSS Korea	+82.2.3446.7651
SPSS Latin America	+1.312.651.3539
SPSS Malaysia	+603.7873.6477
SPSS Mexico	+52.5.682.87.68
SPSS Miami	+1.305.627.5700
SPSS Norway	+47.22.40.20.60
SPSS Polska	+48.12.6369680
SPSS Russia	+7.095.125.0069
SPSS San Bruno	+1.650.794.2692
SPSS Schweiz	+41.1.266.90.30
SPSS Singapore	+65.324.5150
SPSS South Africa	+27.11.807.3189
SPSS South Asia	+91.80.2088069
SPSS Sweden	+46.8.506.105.50
SPSS Taiwan	+886.2.25771100
SPSS Thailand	+66.2.260.7070
SPSS UK	+44.1483.719200