Data Mining, Lecture 6: Cluster Analysis case of categorical data

S. Nõmm

[1]Department of Computer Science, Tallinn University of Technology

March 7, 2016

Introduction

Problems to address:

- Distance computation
- Representative selection
- Density estimation

Above mentioned operations were naturally developed for numeric data.

One may suggest to convert the data categorical data into binary.

Representative based algorithms

- **Centroid of a categorical data set:** (In the case of numerical data, computed by averaging) is a probability histogram of values on each attribute. For each attribute i, and possible value v_j , the histogram value p_{ij} represents the fraction of the number of objects in the cluster for which attribute i takes on value $v_j \Rightarrow$, for a d-dimensional data set, the centroid of a cluster is a set of d different histograms, representing the probability distribution of categorical values of each attribute in the cluster.
- **Similarity:** The simplest of these is match-based similarity. The goal is to determine the similarity between a probability histogram (corresponding to a representative) and a categorical attribute value.

Example

Data	(Color, Shape)
1	(Blue, Square)
2	(Red, Circle)
3	(Green, Cube)
4	(Blue, Cube)
5	(Green, Square)
6	(Red, Circle)
7	(Blue, Square)
8	(Green, Cube)
9	(Blue, Circle)
10	(Green, Cube)

Cluster

Attribute	Histogram	Mode
Color	Blue= 0.4 Green = 0.4 Red = 0.2	Blue or Green
Shape	Cube = 0.4 Square = 0.3 Circle = 0.3	Cube

Mean histogram and modes.

k - Modes and k - Medoids

- k- Modes; representative is also a categorical data record; does not work well with skewed data like market basket data.
- k- Medoids
- Hierarchical Algorithms

Example: ROCK algorithm

ROCK (RObust Clustering using linKs) algorithm is an agglomerative approach in which the clusters are merged on the basis of a similarity criterion.

- Based on Shared Nearest Neighbors.
- Applies the approach to a sample of data.
- Data binarization. (at this point we may talk about transactions).
- Jaccard coefficient is used to define similarity between the sets of transactions.

$$S(T_i, T_j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|}$$

Example: ROCK algorithm

- Two data sets are defined as neighbours if the similarity between them is greater than some threshold θ.
- Similarity leads the graph structure (nodes are data items and links correspond to the neighborhood relations).
- Denote $L(T_i, T_j)$ shared neighbor similarity function. Which is the merging criterion for the agglomerative data algorithms.
- Group link for the clusters $\mathcal{C}_i$ and $\mathcal{C}_j$ is defined as follows:

$$L_{\mathcal{G}} = \sum_{T_u \in \mathcal{C}_i, T_v \in \mathcal{C}_j} L(T_u, T_v).$$

Probabilistic approach

- The main difference from numeric clustering is that the soft assignment process in the E-step, and the parameter estimation process in the M-step will depend on the relevant probability distribution model for the corresponding data type.
- Let the k components of the mixture be denoted by $\mathcal{G}_1, \ldots, \mathcal{G}_k$ the generative process for each point in the data set $\mathcal{D}$ uses the following two steps:
  - Select a mixture component with prior probability α_i, where $i \in \{1, \ldots, k\}$.
  - Generate a data point from the component selected on the previous step.
- The values α_i denote the prior probabilities $P(\mathcal{G}_i)$.
- The main difference from the numerical case is in the mathematical form of the generative model for the m^{th} cluster (or mixture component) $\mathcal{G}_m$, which is now a discrete probability distribution rather than the probability density function used in the numeric case.

Probabilistic approach

- Assume that the j^{th} categorical value of i^{th} attribute is independently generated by mixture component (cluster) m with probability p_{ijm}.
- Consider a data point $\bar{X}$ containing the attribute value indices $j_1, \ldots, j_d$ for its d dimensions.
- The entire set of model parameters is denoted Θ.
- The discrete probability distribution is as follows:

$$g^{m,\Theta}\big(\bar{X}\big) = \prod_{r=1}^{d} p_{ij_r m}.$$

- Posterior probability $P\big(\mathcal{G}_m|\bar{X},\Theta\big)$ may be estimated as follows

$$P\big(\mathcal{G}_m|\bar{X},\Theta\big) = \frac{\alpha_m g^{m,\Theta}\big(\bar{X}\big)}{\sum_{r=1}^{k} \alpha_r g^{r,\Theta}\big(\bar{X}\big)}$$

Probabilistic approach

- (E-step) Posterior probability $P\big(\mathcal{G}_m|\bar{X},\Theta\big)$ may be estimated as follows
$$P\big(\mathcal{G}_m|\bar{X},\Theta\big)= \frac{\alpha_m g^{m,\Theta}\big(\bar{X}\big)}{\sum_{r=1}^{k}\alpha_r g^{r,\Theta}\big(\bar{X}\big)}$$
- (M-step) applies maximum likelihood estimation to the individual components of the mixture to estimate the probability p_{ijm}.
$$p_{ijm} = \frac{w_{ijm}}{\sum_{\bar{X}\in\mathcal{D}} P\big(\mathcal{G}_m|\bar{X},\Theta\big)}$$
where w_{ijm} number of data points in cluster m for which attribute i takes j^{th} possible categorical value.