Clustering Categorical Data

Yi Zhang, Ada Wai-chee Fu, Chun Hing Cai, Pheng Ann Heng

Department of Computer Science and Engineering
The Chinese University of Hong Kong

In this paper we propose two methods to study the problem of clustering categorical data. The first method is based on dynamical system approach. The second method is based on the graph partitioning approach.

Dynamical systems approach for clustering categorical data have been studied by some authors [1]. However, the proposed dynamic algorithm cannot guarantee convergence, so that the execution may get into an infinite loop even for very simple data. We define a new configuration updating algorithm for clustering categorical data sets. Let us consider a relational table with k fields, each of which can assume one of a number of possible values. We prepresent each possible value in each possible field by an abstract node. Let us denote the nodes by $v_i (i = 1, \cdots, m)$. A configuration is an assignment of weight w_i for each node v_i. The new algorithm is defined as follows.

To update the configuration W:
create a temporary configuration W' with
weights $w'_1, ..., w'_m$
for each weight $w_{u_i} \in W$
  {for each tuple $\tau = \{v_{u_1}, v_{u_2}, \cdots, v_{u_k}\}$
   containing v_{u_i} do
   $x_\tau \leftarrow w_{u_1} + \cdots + cw_{u_i} + \cdots + w_{u_k}$.
   $w'_{u_i} \leftarrow \sum_\tau x_\tau$.
  }
  $W \leftarrow W'$
  $W \leftarrow N(W)$

In the algorithm c is a certain positive constant. $N(\cdot)$ is a normalization function, it normalize the weights of all attribute values so that their squares add up to 1. The above algorithm can be mapped into a dynamical system as follows

$$W(k+1) = \frac{QW(k)}{\sqrt{[QW(k)]^T [QW(k)]}}.$$

Theorem 1: *Suppose that all the eigenvalues of Q are positive, then the solution starting from any nonzero points in R^n converges to an unit eigenvector of Q.*

The proof of the theorem can be found in [2]. This theorem guarantees the convergence of the above algorithm by choosing a large enough constant c.

The second approach is based on graph partitioning method. Suppose there are two attributes a and b. A is an attribute value for a and B is an attribute value for b. Let P_A be the fraction of tuples containing A in attribute a. Let P_B be the fraction of tuples containing B in attribute b. If A has no special relationship with B, then we expect that the number of tuples containing both A and B to be $P_A \times P_B$. We use the deviation from this expectation to estimate the relationship between A and B. Hence if P_{AB} is the actual fraction of tuples containing both A and B,

$$linkage_{AB} = \frac{P_{AB} - P_A P_B}{P_A P_B}$$

The algorithm is a variant of the single link clustering method. We scan the database tuples once. With the scanning, we count the number of tuples that contain each attribute value and each pair of attribute values. We can then compute the linkage $linkage_{AB}$. The greater the value of $linkage_{AB}$, the shorter the distance between the values A and B.

Initially we assume each attribute value belong to a unique cluster. We shall try to form bigger clusters by iteratively joining clusters. The joining is done by making use of the linkage of the greatest value in the linkage list. Suppose the linkage is for the pair of values A and B. Then if A and B belong to two different clusters C_1, C_2, we merge the two clusters into one. We repeat the process with the top element remaining in the linkage list. This is repeated until we have exactly two clusters left.

We demonstrated the performance of the proposed algorithms in computation efficiency and also effectiveness by experiments on both real and synthetic data [2].

References

[1] D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamical systems. In *Proceedings of the 24th VLDB Conference, New York, USA*, 1998.

[2] Yi Zhang, Ada W. Fu, C.H. Cai, and Pheng-Ann Heng. Clustering categorical data. *Full paper*, 1999.