Available online at www.sciencedirect.com
SCIENCE DIRECT®
Data & Knowledge Engineering 54 (2005) 189–210
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Knowledge discovery by probabilistic clustering of distributed databases

Sally McClean *, Bryan Scotney, Philip Morrow, Kieran Greer

School of Computing and Information Engineering, University of Ulster, Cromore Road, Coleraine BT52 1SA, Northern Ireland

Received 17 July 2004; accepted 1 December 2004
Available online 23 December 2004

Abstract

Clustering of distributed databases facilitates knowledge discovery through learning of new concepts that characterise common features and differences between datasets. Hence, general patterns can be learned rather than restricting learning to specific databases from which rules may not be generalisable. We cluster databases that hold aggregate count data on categorical attributes that have been classified according to homogeneous or heterogeneous classification schemes. Clustering of datasets is carried out via the probability distributions that describe their respective aggregates. The homogeneous case is straightforward. For heterogeneous data we investigate a number of clustering strategies, of which the most efficient avoid the need to compute a dynamic shared ontology to homogenise the classification schemes prior to clustering.

Keywords: Distributed databases; Probabilistic clustering; Aggregates; Dynamic shared ontology

* Corresponding author. Tel.: +44 2870 324602/1265 324602; fax: +44 2870 324916/1265 324916.
E-mail addresses: si.mcclean@ulster.ac.uk (S. McClean), bw.scotney@ulster.ac.uk (B. Scotney), pj.morrow@ulster.ac.uk (P. Morrow), krc.greer@ulster.ac.uk (K. Greer).

0169-023X/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2004.12.001

1. Background

Clustering of distributed databases enables the learning of new concepts that characterise important common features and differences between datasets. For example, we may have a number of supermarkets belonging to a multinational chain, and each supermarket maintains a database describing their customers. Then we may cluster the databases to learn new high-level concepts that characterise groups of supermarkets. More generally, with increasingly more databases becoming available on the Internet, such an approach affords an opportunity to globalise knowledge discovery and learn general patterns, rather than restricting learning to specific databases from which the rules may not be generalisable.

In this paper we are concerned with clustering databases that hold aggregate count data in the form of *datacubes* on a set of attributes that have been classified according to homogeneous or heterogeneous classification schemes. Such data are often stored in a OLAP style database or Data Warehouses but may also be obtained by pre-processing native databases to provide materialised aggregate views; these may be readily computed from the underlying databases as the result of SQL queries. Such aggregate views are commonly used for summarising information held in very large databases, typically those encountered in data warehousing, large-scale transaction management, and statistical databases. An important special case of native databases that may be summarised in this way is provided by item set data, which stores, for example, binary data on whether a customer bought each item in a given set of possible items in a shopping transaction.

Example 1 (*Heart disease databases* (*Submitted by David Aha to the ML Repository*)). There are four databases each containing 76 attributes, including the decision variable Coronary Heart Disease (CHD), with values 0 (no CHD) to 4 (severe CHD). The databases are respectively: 1. Cleveland Clinic Foundation, 2. Hungarian Institute of Cardiology, 3. University Hospital Zurich and Basel, 4. Long Beach Clinic Foundation.

In general we can cluster on the joint distribution (cross-product) of the decision variable with some of the explanatory variables, e.g., sex, presence or absence of exercise-induced angina. In the example presented in Table 1 we illustrate the approach by clustering the joint distribution of sex with the decision variable (CHD), i.e., we cluster on the vector of proportions in the joint distribution $\{\text{CHD}\} \times \{\text{sex}\}$ for each datacube. The clusters identified are {Cleveland, Hungarian}, {Swiss} and {Long Beach}.

Semantic heterogeneity is a common occurrence in distributed databases, where typically the databases have developed independently. We consider situations where, for a common concept,

Table 1
Clusters: {Cleveland, Hungarian}, {Swiss}, {Long Beach}

Decision variable	Male					Female					Total
	0	1	2	3	4	0	1	2	3	4	
Cleveland	72	9	7	7	2	92	46	29	28	11	303
Hungarian	69	5	1	3	3	119	32	25	25	12	294
Swiss	0	6	3	1	0	8	42	29	29	5	123
Long Beach	3	3	0	0	0	48	53	41	42	10	200