Knowledge Discovery from Semantically Heterogeneous Aggregate Databases Using Model-Based Clustering

Shuai Zhang, Sally McClean, and Bryan Scotney

School of Computing and Information Engineering, University of Ulster, Coleraine, Northern Ireland, UK
{zhang-s1,si.mcclean,bw.scotney}@ulster.ac.uk

Abstract. When distributed databases are developed independently, they may be semantically heterogeneous with respect to data granularity, scheme information and the embedded semantics. However, most traditional distributed knowledge discovery (DKD) methods assume that the distributed databases derive from a single virtual global table, where they share the same semantics and data structures. This data heterogeneity and the underlying semantics bring a considerable challenge for DKD. In this paper, we propose a model-based clustering method for aggregate databases, where the heterogeneous schema structure is due to the heterogeneous classification schema. The underlying semantics can be captured by different clusters. The clustering is carried out via a mixture model, where each component of the mixture corresponds to a different virtual global table. An advantage of our approach is that the algorithm resolves the heterogeneity as part of the clustering process without previously having to homogenise the heterogeneous local schema to a shared schema. Evaluation of the algorithm is carried out using both real and synthetic data. Scalability of the algorithm is tested against the number of databases to be clustered; the number of clusters; and the size of the databases. The relationship between performance and complexity is also evaluated. Our experiments show that this approach has good potential for scalable integration of semantically heterogeneous databases.

Keywords: Model-based clustering, Semantically heterogeneous databases, EM algorithm.

1 Introduction

In fast developing distributed open environments, e.g., the Semantic Web [1], for the same problem domain, distributed databases may be developed independently by different organisations using various ontologies. These databases can be semantically heterogeneous, arising from the use of different terminologies, granularities of data, schemas (conceptualisation) at which objects and their properties are described, and embedded heterogeneous context information [2]. This heterogeneity brings a considerable challenge for distributed knowledge discovery on those databases, for organisations that have common application interests and are willing to cooperate with each other. Most DKD methods in the literature assume that the distributed data are somehow partitioned either horizontally or vertically from a single virtual global table,

R. Cooper and J. Kennedy (Eds.): BNCOD 2007, LNCS 4587, pp. 190–202, 2007.

where all the data have the same statistical distribution, and share the same semantics. However, this assumption does not hold in most practical applications [3]. Distributed data sources contain various underlying semantics due to different backgrounds, environment and purposes when they were developed. A single global view (table) is not sufficient to describe all the distributed data; instead, two or more integrated virtual global tables are needed to capture different data distributions and semantics.

In this paper, we are concerned with heterogeneous databases where the heterogeneity is caused by different classification schemes. Such data are often summarised in Data Warehouses. The summaries may be obtained by pre-processing native databases to provide materialised aggregate views of the information held in very large databases. The objective of our work is to capture different underlying characteristics of these distributed databases, while resolving heterogeneity issues efficiently. We propose a model-based clustering method on the distributed heterogeneous aggregate counts data that are obtained by data summaries. A mixture model is constructed where the databases that are in the same cluster share the same semantics and can be integrated to one virtual global table; and they are different from databases in other clusters that correspond to different virtual global tables. Our approach carries out the integration as part of the clustering process, and the heterogeneity is resolved without previously having to homogenise the heterogeneous local schema to a shared schema. In this way, all the data information available is used for carrying out the clustering, which should lead to better results than methods that are based on data homogenisation. New knowledge can be discovered from the generated global tables, and latent information in the databases is made explicit. The clusters represent different signature profiles of the distributed databases based on proportions (probabilities) of particular values of attributes. For example, for supermarket shopping data from distributed chain-stores, each cluster contains local stores that have similar customer shopping patterns. Each cluster of stores may be of course geographically distributed. The learned clusters contain useful commercial information. When it is required to classify a new instance, only the relevant cluster information is needed instead of the whole data for all the stores. In general, the learned clusters can be used for the construction of Bayesian Belief Networks; alternatively association rules of interest can be extracted [5].

The proposed algorithm evaluation is carried out on both real and synthetic data. Scalability is tested against the number of datasets and the size of the dataset. A clustering complexity measure is designed, and the relationship between the performance (accuracy, computation time) and complexity is evaluated.

The rest of the paper is organised as follows: the data model is briefly introduced, followed by an introduction to the principles of model-based clustering. Clustering of homogeneous data is discussed initially. We then describe our proposed model-based clustering method for heterogeneous data, with an illustrative example. Finally we present algorithm evaluation and conclusions.

2 Terminology and Data Models

Definition 1: An ontology is defined as the Cartesian-product of a number of attributes $A_1,\ldots,A_n$, along with their corresponding schema. The attributes we are concerned with are categorical attributes.