A New Approach for Distributed Density Based Clustering on Grid Platform

Nhien-An Le-Khac, Lamine M. Aouad, and M-Tahar Kechadi

School of Computer Science and Informatics,
University College Dublin, Dublin 4, Ireland
{Nhien-An.Le-Khac,Lamine.Aouad,Tahar.Kechadi}@ucd.ie
http://www.csi.ucd.ie/

Abstract. Many distributed data mining *DDM* tasks such as distributed association rules and distributed classification have been proposed and developed in the last few years. However, only a few research concerns distributed clustering for analysing large, heterogeneous and distributed datasets. This is especially true with distributed density-based clustering although the centralised versions of the technique have been widely used fin different real-world applications. In this paper, we present a new approach for distributed density-based clustering. Our approach is based on two main concepts: the extension of local models created by *DBSCAN* at each node of the system and the aggregation of these local models by using tree based topologies to construct global models. The preliminary evaluation shows that our approach is efficient and flexible and it is appropriate with high density datasets and a moderate difference in dataset distributions among the sites.

Keywords: distributed data mining, distributed clustering, density-based, large dataset, tree topology.

1 Introduction

Today a deluge of data are collected from not only science fields but also industry and commerce fields. Massive amounts of data that are being gathered and stored in different sites. In this context, distributed data mining (*DDM*) techniques have become necessary for analyzing these large and multi-dimensional datasets. Actually, in order to tackle large, graphically distributed, high dimensional, multi-owner, and heterogeneous datasets, some projects have just been started such as Knowledge Grid[2], Grid Miner[1] and ADMIRE[12]. The last project is a new *DDM* framework which is being developed in the Department of Computer Science at University College Dublin. ADMIRE is not only a platform based on *P2P-Grid*[5] infrastructure for implementing *DDM* techniques but also it provides new distributed algorithms for exploring very large and distributed datasets. The first step of the development of these algorithms concern distributed clustering techniques that have few of researches by comparison with distributed association rules and distributed classification.

R. Cooper and J. Kennedy (Eds.): BNCOD 2007, LNCS 4587, pp. 247–258, 2007.

There are two major strands of research into distributed clustering: parallel clustering and distributed clustering. In the first strand, researchers developed parallel versions of the centre-based clustering algorithms. The second strand is based on two principal steps: perform partial analysis on local data at individual sites and then generate a global model by aggregating these local results. Although this later strand is more appropriate for Grid platforms where datasets are often geographically distributed and owned by different organisations, there is more research work in the first strand than in the second[9]. This is especially true with distributed clustering approaches based on density. In this context, recent researches[9][10] have proposed a distributed clustering consisted of two steps: local clustering to build local models and global clustering on these models to build a global model. Global clustering could not scale well when huge amounts of data are available in large-scale networks. In this paper, we propose a new approach of distributed density based clustering. This new approach is composed of the local clustering and the hierarchical aggregation of local models to rebuild a global model.

In our approach, the aggregating process is based on a decentralized model and the local clustering is a density-based. Density-based clustering approaches have been widely used in mining large dataset. Moreover, density based clustering algorithms have been recognized to be powerful and capable of discovering arbitrary shapes of clusters as well as dealing with noise and outliers. There are some density based algorithms such as DenClue[6] and DBSCAN[4]. In this paper, DBSCAN is chosen because it is simple and efficient in very large databases. It requires a minimum domain knowledge to determine input parameters and discover clusters with arbitrary shapes[4]. The rest of this paper is organized as follow: Section 2 deals with background and related projects then we will present and discuss our new distributed density based clustering in section 3. Section 4 presents our preliminary evaluations of this approach. Finally, we conclude on Section 5.

2 Related Works

In spite of a large amount of research conducted in distributed clustering such as [11][18][16], there are very few algorithms proposed in distributed density based clustering. Until now, to the best of our knowledge, there are four approaches in this paradigm that were presented in[17][9][10] and [13]. The former deals with a parallel approach of DBSCAN algorithm. This approach is appropriate for shared memory or distributed shared memory systems. The last three approaches include two main steps: local clustering to create local model and processing these local models to rebuild a global model.

In [9], authors used DBSCAN as a local clustering algorithm. They extended primitive elements of this algorithms such as core points, ϵ, *Minpts* by adding new concepts as specific core points, specific ϵ_{range} to build a local representative at each site. The global model will be rebuilt by executing the DBSCAN algorithm on a set of local representatives with two global values: $Mintps_{global}$ and ϵ_{global}.