Privacy Preserving Clustering for Distributed Homogeneous Gene Expression Data Sets

Xin Li, Georgetown University, USA

ABSTRACT

In this paper, the authors present a new approach to perform principal component analysis (PCA)-based gene clustering on genomic data distributed in multiple sites (horizontal partitions) with privacy protection. This approach allows data providers to collaborate together to identify gene profiles from a global viewpoint, and at the same time, protect the sensitive genomic data from possible privacy leaks. The authors developed a framework for privacy preserving PCA-based gene clustering, which includes two types of participants such as data providers and a trusted central site. Within this mechanism, distributed horizontal partitions of genomic data can be globally clustered with privacy preservation. Compared to results from centralized scenarios, the result generated from distributed partitions achieves 100% accuracy by using this approach. An experiment on a real genomic data set is conducted, and result shows that the proposed framework produces exactly the same cluster formation as that from the centralized data set.

Keywords: Clustering, Gene Profiling, Human Genomic Data, Principal Component Analysis, Privacy Preserving Data Mining

INTRODUCTION

In recent years, new bioinformatics technologies, such as gene expression microarrays, have been widely used to simultaneously identify a huge number of human genomic biomarkers, generate a tremendously large amount of data, dramatically increase the knowledge on human genomic information, and thereafter, significantly improve biomedical research.

A DNA microarray (Wikipedia, 2010), which is the practical realized technology of the Gene Expression (BioChemWeb.org, 2010), is a multiplex technology used in molecular biology. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles (10^{-12} moles) of a specific DNA sequence, known as probes (or reporters). This can be a short section of a gene or other DNA element that is used to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target. Since an array can contain tens of thousands of probes, a microarray...
experiment can accomplish many genetic tests in parallel. Therefore arrays have dramatically accelerated many types of investigation. The microarray data processing pipeline (Hackl, Sanchez Cabo et al., 2004) includes a variety of statistical steps: pre-processing (including background correction, normalization, and summarization), differential analysis which contains raw p-value computation and false discovery rate (FDR) correction, and gene clustering / profiling analysis. Figure 1 shows that microarray experiment process in the lab and Figure 2 illustrates its gene clustering result.

However, these exciting advances do come with an inevitable issue, that is, the richer and richer human genomic data contains privacy sensitive information, such as, genetic markers, diseases, etc., which may further lead to an individual’s race, family, or even identity. Unfortunately, because genomic data does not directly carry individual identity information and it used to be believed that the high-throughput data generated from technologies, such as microarray, is not accurate enough, privacy issues in the human genomic data have not been fully considered as a big issue for quite a while in both biomedical and informatics domains, until a very recent work showed that it was very possible to identify the presence of an individual trace contributor within a series of highly complex genomic mixtures (Homer, Szelinger et al., 2008) under some circumstances. As an immediate response to this new finding, The National Institute of Health (NIH) has agreed to shut down public access not just to individual genotype data but even to aggre-

Figure 1. DNA microarray experiment
Two Stage Capacitated Facility Location Problem: Lagrangian Based Heuristics
www.igi-global.com/chapter/two-stage-capacitated-facility-location/69893?camid=4v1a