Chapter 14
Privacy Preserving Principal Component Analysis Clustering for Distributed Heterogeneous Gene Expression Datasets

Xin Li
Georgetown University Medical Center, USA

ABSTRACT
In this paper, we present approaches to perform principal component analysis (PCA) clustering for distributed heterogeneous genomic datasets with privacy protection. The approaches allow 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. We then further develop a framework for privacy preserving PCA-based gene clustering, which includes two types of participants: data providers and a trusted central site (TCS). Two different methodologies are employed: Collective PCA (C-PCA) and Repeating PCA (R-PCA). The C-PCA requires local sites to transmit a sample of original data to the TCS and can be applied to any heterogeneous datasets. The R-PCA approach requires all local sites have the same or similar number of columns, but releases no original data. Experiments on five independent genomic datasets show that both C-PCA and R-PCA approaches maintain very good accuracy compared with the centralized scenario.

INTRODUCTION
In recent years, new bioinformatics technologies, such as gene expression microarray, genome-wide association study, proteomics, and metabolomics, have been widely used to simultaneously identify a huge number of human genomic/genetic biomarkers, generate a tremendously large amount of data, and dramatically increase the knowledge on human genomic/genetic information, thus significantly improving biomedical research. However, these exciting advances in bioinformatics do come with a drawback: the increasingly richer human genomic/genetic data contains sensitive private information, such as genetic markers, diseases,

DOI: 10.4018/978-1-4666-2653-9.ch014
Privacy Preserving Principal Component Analysis Clustering etc., which may further lead to the discovery of the individual’s race, family, or even identity. Therefore, privacy is an important issue when dealing with bioinformatics data. This is further exacerbated when multiple data providers try to collaborate with each other.

**Gene Expression and DNA Microarray**

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, Sturn, Wolkenhauer, & Trajanoski, 2004) includes a variety of statistical steps: pre-processing (including background correction, normalization, and summarization),
Related Content

Statistical Models in Bioinformatics
Stelios Zimeras and Anastasia N. Kastania (2010). Bio computation and Biomedical Informatics: Case Studies and Applications (pp. 128-142).
www.igi-global.com/chapter/statistical-models-bioinformatics/39608?camid=4v1a

Discriminative Subgraph Mining for Protein Classification
Ning Jin, Calvin Young and Wei Wang (2010). International Journal of Knowledge Discovery in Bioinformatics (pp. 36-52).
www.igi-global.com/article/discriminative-subgraph-mining-protein-classification/47095?camid=4v1a

Information Therapy (Ix) Service and Patients’ Preference
www.igi-global.com/chapter/information-therapy-service-patients-preference/73074?camid=4v1a

Data Mining Approach for the Early Risk Assessment of Gestational Diabetes Mellitus
www.igi-global.com/article/data-mining-approach-for-the-early-risk-assessment-of-gestational-diabetes-mellitus/202360?camid=4v1a