Guest Editorial Preface

Special Issue on Mining Big Biomedical Data

Samah Jamal Fodeh, Yale University, New Haven, CT, USA

INTRODUCTION

The exponential and continuous growth of big biomedical data stimulated the need to developing advanced and sophisticated big data analytics. In response to this need, this special issue was started. Researchers with different focuses were invited to participate in this issue including those working on problems concerning data clustering, classification, visualization, information retrieval and data integration. The call for papers encouraged researchers to mine different types of Big Data including genomic, proteomic phenotypic, molecular (including –omics), physiological, anatomical, clinical, behavioral, environmental, social media, and many other types of biological and biomedical data.

International Journal of Knowledge Discovery in Bioinformatics (IJKDB) is particularly interested in publishing methodological reviews on topics that introduce methodological innovations. Containing articles on topics such as systems biology, protein structure, gene expression, and biological data integration. IJKDB journal presents a cross-disciplinary approach to the field useful for researchers, practitioners, academicians, mathematicians, statisticians, and computer scientists involved in the many facets of bioinformatics.

Manuscripts targeting many research problems were submitted to the Mining Big Biomedical Data special issue. Four submissions were selected to be included in the final issue. In this preface, I provide a snapshot of the research described in the special issue. Two manuscripts are cancer studies concerned with utilizing advanced machine learning techniques and the other two manuscripts utilized MapReduce to handle gene expression data and big Web data.

To handle the inherent problem of high dimensionality and small sample size of gene expression microarray data, Dash and Patra (2016) proposed a new hybrid approach that combines feature selection with ensemble learning for genetic diagnosis of cancer. Specifically, they used fuzzy-rough feature selection model and an adaptive neural net ensemble learning algorithm. While the fuzzy-rough method deals with uncertainty and impreciseness of real valued gene expression data, ensemble neural networks improves learning compared to single classifiers and existing ensemble methods. The experimental analysis showed that their model has achieved highest averaged generalization ability compared to its counterparts and established an acceptable level of diversity among the base learners for all the benchmark datasets. In another study, Jananee and Nedunchelian (2016) utilized association rule mining (Agrawal, Rakesh, Imieliński, & Swami, 1993) to detect genes causing breast cancer. First association rules are generated and the log likelihood of each rule is computed using the Baum-Welch process (an iterative procedure that helps to find the unknown parameters of Hidden Markov Model (HMM) by calculating the actual and emission probabilities). The geneses are then

clustered into three groups: high, medium low based on their log likelihood. The genes in the high valued cluster are considered to be breast cancer causing genes.

Gowri and Rathipriya (2016) used MapReduce to mine local patterns in gene expression data via bi-clustering. MapReduce is a framework used to perform computations on large dataset and can cope with scalability problems (Dean & Ghemawat, 2008). Using the bi-clustering approach introduced in (Rathipriya, Thangavel, & Bagyamani, 2011), they proposed to simultaneously select rows (genes) and columns (conditions) in yeast (saccharomyces cerevisiae) gene expression data. The identified biclusters preserved locally defined degree of homogeneity between genes and conditions of species. The beneficial outcome of the proposed method is the high correlation between the genes and conditions of a species or interspecies. It also outperforms the existing traditional bi-clustering methods of gene expression data. Although this approach was developed and applied to gene data, Rathipriya (2011) took another step forward as he modified the algorithm and applied it to Web Data. We include his manuscript in this special issue as well. Similarly, he used MapReduce for bi-clustering but he contributed the correlation based similarity measure called ACV used to capture pattern-based closeness of web users. His goal was to identify users with similar web usage profiles. This measure could be useful for E-Commerce applications such as marketing, advertising, recommendation, as it provides valuable information about users' interests.

The manuscripts published in this special issue is an effort to address some of the obstacles related to the high dimensionality and patterns discovery in Big Data. However, mining big biomedical data remains an open and challenging research problem.

Samah Jamal Fodeh
Guest Editor
IJKDB

REFERENCES

Dash, S., & Patra, B. (2016). Genetic diagnosis of cancer by evolutionary fuzzy-rough based neural-netowrk ensemble. *International Journal of Knowledge Discovery in Bioinformatics*, *6*(2).

S., J. & R., N. (2016). Detection of breast cancer by the identification of circulating tumor cells using association mining. *International Journal of Knowledge Discovery in Bioinformatics, 6*(2).

Rakesh, A., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *SIGMOD Record*, *22*(2).

R., G. & R., R. (in press). Local optima avoidance in GA biclustering using MapReduce. *International Journal of Knowledge Discovery in Bioinformatics, 6*(1).

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107–113. doi:10.1145/1327452.1327492

Rathipriya, R., Thangavel, K., & Bagyamani, J. (2011). Evolutionary Biclustering of Clickstream Data. *International Journal of Computer Science Issues*, *8*, 32–38.

R., R. (in press). A novel evolutionary biclustering approach using MapReduce (EBC-MR). *International Journal of Knowledge Discovery in Bioinformatics, 6*(1).