Clinico-genomic Decision Support System for Precision Diagnostics and Management

Anuja Kench[1], Vandana P. Janeja, PhD[1], Amanda Niskar, DrPH[2], Yelena Yesha, PhD[1], Napthali Rishe, PhD[3]

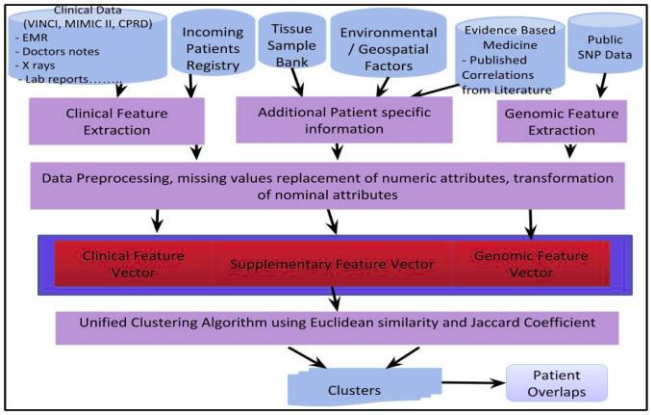
[1]University of Maryland Baltimore County, Baltimore, USA,

[2]Arthritis Foundation

[3]Florida International University, Florida, USA

Introduction

The focus of this work is clinico-genomic decision support system for better diagnosing and managing chronic diseases. We propose a clinical decision support system for early diagnosis of chronic diseases(such as arthritis, Diabetes, Cardio Vascular Disease) and better disease management. We propose a generalizable framework, which can accommodate several distinct features coming into the system from clinical, genomic, environmental domains among others. We utilize this framework to discover similar patients and overlaps among patients in a set of these features which has several applications such as : (a) better cohort discovery for clinical trials, (b) better disease management by studying peer group of patients with similar diagnosis and better prognosis, (c) early disease diagnosis by studying similar features in patients with existing diagnosis



Approach: The overall approach, as shown in the figure, looks at clinical data and performs clinical record extraction. The clinical data consist of the structured medical record, unstructured doctor's notes, X-rays, lab data among other data components. Some studies have looked at extracting and integrating structured and unstructured data[BM15, G14]. However, our work focuses on integrating multiple types of data from multiple sources with different types of structure. Data is extracted from relational and non-relational forms. Similarly, genomic data is extracted from a large learning database and matched with patient specific genomic data. Databases like VINCI [VA2015] and dbGap[NCBI2015] contain such heterogeneous patients records that can be accessed under agreement. The genomic data is in the form of SNPs extracted from large databases [SW01] and prior clinical studies and patient specific genomic data. Data transformation and preprocessing is performed as new variables are introduced at different stages of the analysis.

The clustering of continuous variables and the categorical variables from clinical and genomic data are combined using a novel algorithm to form clinico-genomic clusters. Data within the clusters is evaluated to find partial or exact patient overlaps. The partial overlaps, especially overlaps above a certain threshold may indicate a high level of similarity in majority of attributes but not in some attributes, which may be of interest to study especially if the overlap is in key clinical variables but not in genomic variables and vice versa. Patients with high overlaps may also be selected as a cohort for clinical trials. The level of similarity can be calibrated to identify various types of overlaps. In addition, multiple diseases can be studied together, to identify potential overlaps between patients from disease groups where no known overlaps exist. This can lead to the discovery of potential links between diseases with no known clinical connections and potentially lead to novel research.

References

[BM15]Bullard J, Murde R, Yu Q. Inference from Structured and Unstructured Electronic Medical Data for Dementia Detection. 14th INFORMS Computing Society Conference, pp. 236–244, 2015.

[NCBI15] http://www.ncbi.nlm.nih.gov/gap Database of Genotypes and Phenotypes (dbGaP), accessed July 3, 2015

[SW01] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation}, Nucleic Acids Res. 2001 Jan 1;29(1):308-11.

[G14] Ghiasvand O. Disease Name Extraction from Clinical Text Using Conditional Random Fields. May 2014.

[VA2015] VA Informatics and Computing Infrastructure (VINCI) http://www.hsrd.research.va.gov/for_researchers/vinci/, Last accessed July 3, 2015.