Information Refining: Improving the Quality of Information Mined from Heterogeneous and High-Dimensional Time Series

Fatih Altiparmak[1], Ozgur Ozturk[1] , Selnur Erdal[1], Hakan Ferhatosmanoglu[1], Donald C. Trost[2]
[1]The Ohio State University, Columbus, OH [2]Pfizer Global Research and Development

Challenges in Mining Heterogeneous, Asynchronous Time Series

Decreasing price of obtaining data w/ technology
⇒ data abundant

Opportunity: Cross validation information from different sources

Difficulty: Data Incompatibility

- Conventional Data Mining (DM) techniques not fit for heterogeneous & high-dimensional time series

Challenges Faced both in Clinical Trials and Microarray High-dimensionality, Heterogeneity, non-uniformity???, Insufficient length, Unequal interval sizes (variable sampling???), Different lengths, Asynchronicity???, Diverse data sources, Varying sensitivity with source, Noise

- Brute Force DM compared with our method

Global mining of data causes inaccuracies even with extensive preprocessing

Results had little meaning

Heterogeneity and incompleteness of data

Difficulty to interpret such results

Our Two Step INFORMATION REFINING Method

First Step

- Apply DM over homogeneous subsets of data, gather information

Second Step

- Refine Information by identifying common or distinct patterns over it

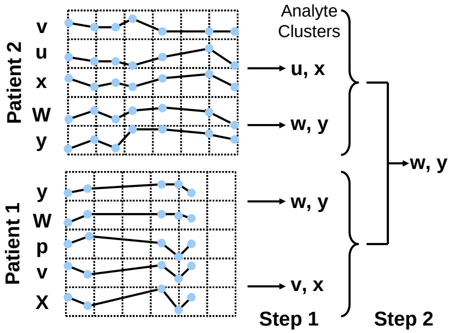
Our Novel Distance Metrics

- Slope Wise Comparison (SWC)
  - Trends matched (increasing or decreasing)
- Qualitative Metric (non-linear correlations)
  - Uses a local distance metric (SWC was used)
  - Local Distance metric must be capable of comparing relationship of two points (a pair) of one series with that of two points of another series
  - Captures the similarity between patterns of changes of time series, regardless of whether the nature of the dependence between them is linear or non-linear.

Case Study 1: Pharmaceutical Clinical Trials

Clinical Trial: A clinical trial is a research study to answer specific questions about vaccines or new therapies. Clinical trials are used to determine whether new drugs or treatments are both safe and effective. In these trials, patients are assigned a treatment or a placebo and measurements for certain analytes (blood ingredients) are taken at intervals. These measurements can be represented as a time series for each analyte.

Information Refining on Clinical Trials



Information Refining Depicted on a Hypothetical Run

Preprocessing

Find significant and clean subsets of data.

e.g. Most appropriate Analytes and Patients to make accurate experiments

-26 (of 43) analytes and 152 patients-

Step 1: Mine the data within clean subsets

Analytes are clustered for each patient

K-Medoid Clustering with 5 different metrics

Output: analyte clusters for each patient

Step2:Refine information (Detect Related)

Input : Analyte clusters for each patient

Find the frequently co-occurring analytes

Merge the analyte sets using

- Support Test
- Confidence Test

Output: Strongly related analyte sets (used in redundancy elimination.)

Refining the Information

Basic Definitions

Support: number of clusters that contain all the members of an analyte-set

Confidence of Association rule X ⇒ Y: Support(X ∪ Y) / Support(X)

Lift (Correlations) of Association rule X ⇒ Y: Support(X ∪ Y) / Support(X)*Support(Y)

To get the strongly related analyte sets of size k,

- **generate candidate sets from the sets of size (k-1)**
- **prune ones that don’t pass support and confidence test**

For example: **{1,2},{1,3},{2,3}** exists ⇒ **{1,2,3}** is a candidate set

```
  IF      Support({1,2,3} > supportLimit
          & Confidence({1,2}, {1,2,3}) > confidenceLimit
          & Confidence({1,3}, {1,2,3}) > confidenceLimit
          & Confidence({2,3}, {1,2,3}) > confidenceLimit
THEN      {1, 2, 3}   is a strongly related analyte set.
```

Findings 1: Strongly Related Analyte Sets Result of Ensemble Algorithm:

Group Name	Group Analytes
Transporter	Hemoglobin, Hematocrit, RBC count
Acute Infection	WBC Count, Neutrophils, Neutrophils (abs)
Serum Protein	Total Protein, Albumin, Globulin, **Calcium**
Liver	SGOT(AST), SGOT(ALT), LDH

Alternative Approach that Finds **Un**related

Run the Algorithm on the Dual of Support values
Total number of patients - support

- **Output:** Selected Features: Global Panels

Feature Selection: Identifying a Global Panel

- A panel of analytes that effectively models the human health

A subset representing all 43 analytes

- Decision support to choose representative(s) from each group of analytes

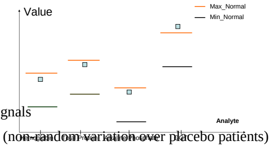
An analyte will be a representative of a panel if it is in a global panel.

Group Name	*Acute Infection*	*Transporter*	*Serum Protein*	*Liver*
Representation frequency	**100%**	**91%**	**87%**	**98%**
Correlation Coefficient	87	100	80	93
Qualitative	100	97	69	100
DTW-Euc	100	100	100	100
DTW-SWC	100	100		100
Euclidian	100	68	98	59

Goals of Pfizer Project

Safety Detection

- Early identification of abnormal individuals to detect safety problems
- Dynamic and multi-dimensional monitoring rules
- Prediction of biomarkers
- Classification of changes
- Current method: Simple univariate normal boundaries:



- We need
  - Multi-variate signals
  - Trajectories??? (non-random variation over placebo patients)
  - Detection of change in correlation of analytes over time

Modeling of health state given clinical measurements

- Healthy vs. Diseased
- Change in health state
- Model the state with less # of analytes?
- How to model the analytes?
  - Feature selection – which analytes are necessary to model a certain health state/disease
  - Global panel of analytes that best represents the overall information in the data
  - Clusters of analytes that represent different groups of biological panels

Acknowledgements

Pfizer???
Children’s Hosp???
BAALC group???

References

- “*Information Mining over Heterogeneous and High Dimensional Time Series Data in Clinical Trials Databases*”, Altiparmak F., Ferhatosmanoglu H., Erdal S., Trost C., IEEE Transactions on Information Technology in Biomedicine (TITB)
- “*Similarity Based Analysis of Microarray Time-Series Data*”, Altiparmak F., Erdal S., Ozturk O., Ferhatosmanoglu H. (Submitted to TITB)

Case Study 2: Haemophilus Influenza Microarray Data

Microarray Technology: A new way of studying how thousands of genes interact with each other and how a cell's regulatory networks control vast batteries of genes simultaneously. The method uses tiny droplets containing functional DNA located as a precise grid on glass slides. Fluorescent labeled DNA probes from the cell being studied are allowed to bind to these complementary DNA strands. Brightness of each fluorescent dot, measured with a scanner, reveals how much of a specific DNA fragment is present, an indicator of how active it is.

Microarray Data

- Usually time series data
- Each series shows change in the expression levels of corresponding gene
  - Measured as density of the gene products existing in cell