Feature Mining For Prediction of Degree of Liver Fibrosis

Benjamin Mayer, Huzefa Rangwala, Rohit Gupta, Jaideep Srivastava, George Karypis, Vipin Kumar, Piet C de Groen*
Department of Computer Science and Engineering, University of Minnesota (200 Union Street SE, Minneapolis MN 55455 USA)
*Department of Health Sciences Research, Mayo Clinic (200 First Street SW, Rochester, MN 55905 USA)
e-coords: rohit@cs.umn.edu

Project Scope

- Build a classification model to predict liver fibrosis stage based on combinations of common laboratory tests, obtained over time during routine patient care
- Find unknown novel patterns in a large clinical data set that may help physicians understand disease co-occurrence and thereby improving treatment procedures
- Identify temporal patterns between hepatitis B and hepatitis C
- Evaluate whether interferon therapy was effective in patients with hepatitis

Problem Characterization

- Temporal
  - Difficult to handle discrete time points as well as varied range of values across the laboratory test
- Sparse
  - Missing values imply the need to use missing data imputation techniques or other ways suitable for this domain
- Noisy
  - Need data cleaning techniques, outlier detection
- Irregular
  - Inability to do direct comparisons between patient test values
- Domain Knowledge
  - Need expertise to understand data, evaluate results, etc.

Problem Definition

- Classification Problem Definition:
  - Prediction of the fibrosis stage of patients liver by evaluating a combination of features obtained (direct: laboratory test values; and derived: observed trends)
  - Class labels i.e. fibrosis stage (f0 – f4) for the training data are obtained from patients liver biopsy pathology report(s)
- Dataset Construction:
  - Only laboratory test values that precede the date of liver biopsy or liver transplant (Whichever is earlier) are used
  - Only those laboratory tests are selected as features for the classification model that shows high (defined by setting a threshold) correlation with the class labels in the training dataset

Feature Extraction

- Computing the correlation between the lab test values and fibrosis stage from the processed data set
  - Result: 11 laboratory tests identified as significant (these laboratory tests were found to present in many patients, hence making a dense dataset)
- Literature survey and physician expertise suggested the use of a subset of laboratory tests
  - Result: 8 commonly used laboratory tests were present in both hepatitis B and C data sets, of which domain expert identified 5 laboratory tests as likely to be important for predicting the course of hepatitis

K-Nearest Neighbor

- A technique that looks at the k points nearest to itself
- Current point is classified based on the count of the neighbors
- Results using 20 nearest neighbors for classification represented as a confusion matrix

For such a medical problem it would be ideal to get the predictions as close to the diagonal. There is a mixing of classes as shown by the red marked region. (Note: There is a mixing in classes especially in stage 1 and stage 2 in fibro test shown below)

PREDICTION

LABELS	f0	f1	f2	f3	f4
f0	0	12	0	0	0
f1	0	165	15	3	4
f2	0	70	20	6	4
f3	0	36	8	7	12
f4	0	30	10	5	19

Confusion Matrix Using K-Nearest Neighbor Approach for K = 20.

Kernel Metrics

- Features Used

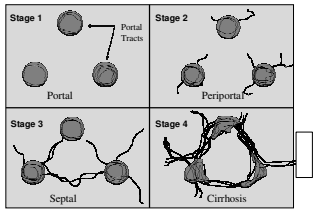
  Dynamic time warping as a direct similarity measure between various patients

  Use of the average of last n test points seen before liver biopsy
- Results (Preliminary)

  Area under the ROC Curve used to evaluate the performance shows that there is promise in this technique, provided
  - the classes are defined well
  - features are combined together

  The best performance value: 0.66 ROC, which is a good estimate given the nature of the problem

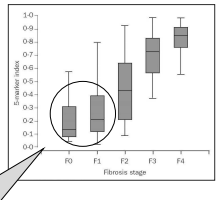
Stages of Fibrosis in Chronic Hepatitis



Slide provided by Dr. John Gross, Mayo Clinic

Laboratory Based Test for Hepatic Fibrosis

FibroTest ®
Note extensive overlap between stages when 5 specific markers are used.
3 of 5 markers are not commonly used tests



Overlap in prediction: similar to K-nearest neighbor results, however this seems to do better for sever classes

Conclusions

- Temporal Evolution of Lab-test Values
  - Trends in the laboratory test values should be considered rather than just the absolute values. The idea is to construct new features that captures the trend in the value of the laboratory test considered till liver biopsy or liver transplant
- Increasing the time span
  - It is believed that increasing the time span of the laboratory test time series would improve the classification results
- Separating hepatitis B and C
  - Hepatitis B cases may have a different pattern in the temporal nature of the laboratory tests compared to Hepatitis C cases
- Integrating phenotypic data (clinical records) with genotypic and other biological data would potentially improve classification results
- Extraction of SNOWMED codes from available pathology reports might help define reference fibrosis classes in an accurate way

Acknowledgements

This work has been supported by DTC grant 1224-04. Computational resources for this work were provided by the Minnesota Supercomputing Institute and Mayo Clinic. Authors also acknowledge Mayo Clinic for providing access to a part of Mayo Clinic Life Sciences System (MCLSS) data.

Association Analysis

- Data: Diagnosis history (ICD9 Codes) of 10,000 patients from Mayo Clinic Life Sciences System (MCLSS) – a large clinical database
- Association rule mining using Apriori Algorithm (http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html) is explored to extract meaningful, interesting and unknown diagnosis patterns across all the patients with one or other liver disease
- Hyperclique pattern mining is also used to extract patterns that involve diagnosis codes with varying support levels (http://www-users.cs.umn.edu/~huix/publication/icdm03.pdf)
- The idea behind extracting such association patterns is to have further insight in the disease co-occurrence and disease progression

Sample Association Rules

- Cirrhosis:
  [HYPERTENSION NOS] [HPT C ACUTE WO HPAT COMA] [CHRNC HPT C WO HPAT COMA] -> [CIRRHOSIS OF LIVER NOS]
  [SCREENING-PULMONARY TB] [LIVER TRANSPLANT STATUS] [ASCITES] -> [CIRRHOSIS OF LIVER NOS]
  [GASTRODUODENAL DIS NEC] [ASCITES] [PORTAL HYPERTENSION] -> [CIRRHOSIS OF LIVER NOS]
- Hepatic Coma:
  [OTH SEQUELA, CHR LIV DIS] [DEBILITY NOS] [PORTAL HYPERTENSION] [CIRRHOSIS OF LIVER NOS] -> [HEPATIC COMA]
  [OTH SEQUELA, CHR LIV DIS] [HYPERTENSION NOS] [ASCITES] [PORTAL HYPERTENSION] -> [HEPATIC COMA]
  [HYPOSMOLALITY] [HYPERPOTASSEMIA] [ASCITES] -> [HEPATIC COMA]
- Portal Hypertension:
  [VARICES OF OTHER SITES] [ESOPH VARICE OTH DIS NOS] [ESOPH VARICES W/O BLEED] [CIRRHOSIS OF LIVER NOS] -> [PORTAL HYPERTENSION]
  [ESOPH VARICE OTH DIS NOS] [GASTRODUODENAL DIS NEC] [HYPERTENSION NOS] [CIRRHOSIS OF LIVER NOS] -> [PORTAL HYPERTENSION]
  [OTH SEQUELA, CHR LIV DIS] [ESOPH VARICE OTH DIS NOS] [GASTRODUODENAL DIS NEC] -> [PORTAL HYPERTENSION]
- Alcoholic Cirrhosis:
  [ALCOH DEP NEC/NOS-UNSPEC] [OTH SPCF PREOP EXAM] [ASCITES] -> [ALCOHOL CIRRHOSIS LIVER]
  [ALCOH DEP NEC/NOS-UNSPEC] [PORTAL HYPERTENSION] [HYPERTENSION NOS] -> [ALCOHOL CIRRHOSIS LIVER]
  [ALCOH DEP NEC/NOS-UNSPEC] [ASCITES] [PORTAL HYPERTENSION] [CIRRHOSIS OF LIVER NOS] -> [ALCOHOL CIRRHOSIS LIVER]
- Biliary Cirrhosis:
  [SCREEN MAL NEOP-CERVIX] [SCREEN MAMMOGRAM NEC] [OSTEOPOROSIS NOS] [HYPERTENSION NOS] -> [BILIARY CIRRHOSIS]
  [CHOLANGITIS] [BONE & CARTILAGE DIS NOS] -> [BILIARY CIRRHOSIS]
  [BREAST DISORDERS NEC] [SCREEN MAL NEOP-CERVIX] [SCREEN MAMMOGRAM NEC] -> [BILIARY CIRRHOSIS]

Note: All the rules above have Support > 3 % and Confidence > 90 %

Insights

- Some of the extracted association patterns that have the following as rule consequent are shown in the above slide: Cirrhosis, Hepatic Coma, Portal Hypertension, Alcoholic Cirrhosis, Biliary Cirrhosis
- Knowledge/Insight gained from association patterns (especially those helpful in understanding disease progression) can be used to take necessary preventive measures and improvement in the treatment procedures
- Patterns extracted were able to capture known medical symptoms. For e.g.
  - Scar liver tissue blocks normal blood circulation through the liver, this blood backs up, leading to increased pressure within the vein (portal hypertension)
  - Blocked circulation in portal vein may cause blood to enter blood vessels of stomach and esophagus, which eventually cause massive bleeding
  - Liver damaged by cirrhosis has trouble removing toxins from the body. The buildup of toxin Ammonia can damage the brain, leading to changes in the mental state and thereby causing Coma

References

- Benjamin Mayer, Huzefa Rangwala, Rohit Gupta, Jaideep Srivastava, George Karypis, Vipin Kumar, Piet de Groen, "Feature Mining for prediction of degree of liver fibrosis", AMIA 2005
- Shahar, Y. and Musen, M. A., "Knowledge-Based Temporal Abstraction in Clinical Domains", Artificial Intelligence in Medicine, 8(1996), 267—298
- Bellazzi, R., Larizza, C., Magni, P., Monntani, S., and Stefanelli, M., "Intelligent Analysis of Clinic Time Series: An Application in the Diabetes Mellitus Domain", Artificial Intelligence in Medicine 20 (2000), 37—57
- Tu Bao Ho, Trong Dung Nguyen, Saori Kawasaki, Si Quang Le, Dung Duc Nguyen, Hideto Yokoi, Katsuhiko Takabayashi, "Mining hepatitis data with temporal abstraction", KDD 2003
- Franky Kin-Pong Chan, Ada Wai-chee Fu, Clement Yu, "Haar Wavelets for Efficient Similarity Search of Time-Series: With and Without Time Warping", TKDE, 2003 (Vol. 15, No. 3), pp. 686-705
- Chiba data set: http://lisp.vse.cz/challenge/ecmlpkdd2004/