Data Mining in the Medical Domain: A Review of some Current Work at The University of Liverpool

Frans Coenen
Department of Computer Science
The University of Liverpool
coenen@liverpool.ac.uk

http://www.csc.liv.ac.uk/~frans/Seminars/lmi12010-11-30.pdf

Presentation Overview

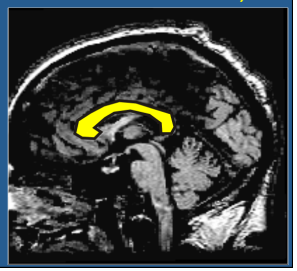
- Data Mining.
- Current work:
  - MRI scan mining.
  - AMD identification.
- Future Directions.
- Summary and conclusions.

What is Data Mining?

- Data mining is the non-trivial discovery of interesting knowledge from data beyond that which can be obtained by simple database querying.
- It can be thought of as "in-depth" data analysis.
- Remit includes: clustering, classification/ categorisation and pattern identification.
- The data mining community is very good at mining tabular data; we wish to apply our techniques to much more challenging data (such as image sets).

Current work (1): MRI scan mining

- Classification of MRI scan data, for medical research purposes, according to a particular feature in these scans called the Corpus Callosum (CC)
- The conjecture is that the shape and size of the CC serves to distinguish; for example, musicians and non-musicians, and handedness.
- It is also suspected that the CC shape and size plays a role in the identification of epilepsy, schitsophrenia, autism, etc.

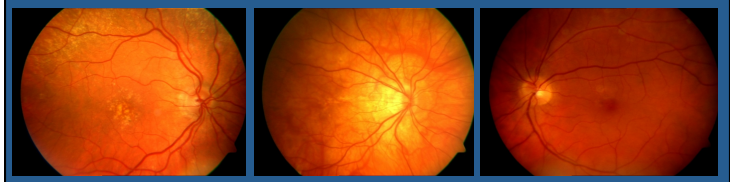


MRI Scan Mining

- Example of image mining according to a particular feature that is known to be present across an image set.
- Involves: feature identification/capture, registration and feature representation.
- Techniques developed to date:
  - **Tree representation (particularly quad-trees)**. Many different ways of "tesselating" images. Tree representations allow for the application of frequent sub-tree mining techniques.
  - **Time series representation**. Many different methods for defining a feature in terms of a curve (radial intersection, moments, Hough transforms). Allows for the application of time series analysis techniques. We have actually coupled this with Case Base Reasoning (CBR).

Current Work (2): AMD identification

- Aim is to provide screening support for the early diagnosis of Age-related Macular Degeneration AMD.
- A common (standard) mechanism for doing this is by identifying "drusen" in retina scans.



AMD Identification

- Example of image mining according to the presence/ absence of a particular type feature in the image set.
- Involves: image registration and representation.
- Techniques developed to date:
  - **Histograms**. Many techniques for generating the desired histograms (which can then be represented as time series, to which we can apply our time series analysis techniques).
  - **Graph representations**. Involves the hierarchical decomposition of an image into a tree structure.
  - **Tabular representations**. Involves segmentation of the images to identify the features of interest.

Further Directions

- A general challenge in the context of image mining is 3-D image mining (image mining applied to volumetric data).
- Issues:
  - Entails a significant computational overhead.
  - Requires appropriate data structures that serve to minimise the identified computational overhead while at the same time supporting appropriate analysis.
- Potential solution:
  - We (Liverpool), in the context of our MRI and AMD work, have developed a sequence of 2-D image mining techniques (tree and graph, time series, histogram and tabular based approaches) and in the process have acquired considerable expertise.
  - The suggestion is that we can extend these approaches to address 3-D (volumetric) image mining.

Applications

- MRI scan data comes in a sequence of slices, currently we are only considering the midsagittal slice. Analysing brain scan features in 3-D may lead to a significant discoveries.

- We also have retina volumetric data obtained using Spectral Domain – Optical Coherence Tomography (OCT). Analysis of this volumetric data may also prove fruitful.

Summary and Conclusions

- We have experience of 2-D image mining in two contexts:
  - i) Image classification according to a specific feature that is known to exist across a given image set, and
  - ii) Image classification according to the presence/absence of a specific feature.
- We would like to extend this experience to the mining of volumetric data.
- We have some ideas on how this might be achieved, and have identified some applications where this might be of benefit.