Digital Pathology Diagnosis Assistance System

Mitotic Figure Recognition: Agreement Among Pathologists and Computerized Detector

3 August 2011
Outline

- Importance of the task
- Inter-observer agreement
- Machine recognition
Grading breast tissue

- Systems for grading invasive breast tumors
  - Greenough (1925) – 7 variables
  - Bloom-Richardson (1957) – 3 variables
  - Nottingham-Bloom-Richardson (1991) – 3 more stable variables

- Original purpose of breast cancer grading [Bloom-Richardson 1957]
  - **Locality**: Higher grades are more likely to have metastasized
  - **Prognosis**: Higher grades have shorter mean survivals

- Grading results inform the selection of high-risk treatments
Cohen’s Kappa: measure of agreement of a pair of observers

Perfect agreement would give kappa = 1

Random guesses with the same answer distributions would give kappa = 0

“Slight agreement”: 0-0.2
“Fair agreement”: 0.2-0.4
“Moderate agreement”: 0.4-0.6
“Good agreement”: 0.6-0.8
“Almost perfect agreement”: 0.8-1
Meyer 2005] investigated inter-observer agreement on components of the Nottingham-Bloom-Richardson grade

- Groups of 5-7 pathologists, each group examining 10-23 cases
- Tissues from Comprehensive Breast Cancer Tissue Resource (CBCTR)

- Overall grade: kappa = .50 - .59
- Tubularity: kappa = .57 - .83
- Pleomorphism: kappa = .27 - .50
- Mitotic grade: kappa = .45 - .67
NEC e-Pathologist Project

- Provide diagnosis support for anatomical pathology
- Gastric, breast, prostate modules in development
- Gastric system for quality control in use in Japan
- Breast: make grading more consistent and more efficient
- Target the least two stable parts of Bloom-Richardson grade
  - **Pleomorphism** – make the grades reflect actual statistics
  - **Mitosis** – apply consistent judgments
- Operate on **hematoxylin & eosin**-stained tissue
Determining the Mitotic Grade

**Bloom-Richardson**
- Count mitosis plus hyperchromatic nuclei
- Vague cutoffs

**Nottingham-Bloom-Richardson** [Elston & Ellis 1991]
- Do not count hyperchromatic nuclei, apoptotic nuclei, or pyknosis
- Do not count prophase, because agreement is low
- Count mitotic nuclei in 10 high-power fields (25X or 40X)
- Fields should be at the periphery, where active growth most likely
- Grading cutoffs depending on HPF size
“Mitotic Impression” (MIMP):
30-second estimate of mitotic grade in 10 HPF
“Eyeball it”

“Mitotic Activity Index” (WHO-MAI):
2-3 minutes careful count

[Jannink 1995]: 10-15 minute procedure, detailed examination of individual nuclei

Between MIMP and WHO-MAI, \( \kappa = .41 \) [Skaland 2008]
433 node-negative non-metastatic invasive ductal breast cancer patients
MIMP is prognostic with \( P=.0003 \)
MAI is prognostic with \( P < .000001 \)
Baak’s Definition of Mitotic Figures

*Hum. Path. 21:683-685 (1990)*

- Nuclear membrane must be absent (=> beyond prophase)
- Clear, hairy extensions of nuclear material (condensed chromosomes) must be present, either clumped (beginning metaphase), in a plane (metaphase/anaphase), or in separate chromosomal aggregates (telophase)
- Cytoplasm may be enlarged
- But NOT:
  - Small, regular instead of hairy extensions with an empty central zone
  - Separate dark round nuclear clumps
  - Orange-colored cytoplasm
One participant’s decision tree…
What about figure-level agreement?

Previous studies are about interobserver differences in grade. But what about agreement on individual figures?

Largest study known to date: [Meyer 2005] – 43 potential mitotic figures, 7 observers
  Average pairwise Kappa = .38

Our study – 4,204 potential mitotic figures
  Three pathologists, actively signing out breast pathology
  Selected to be mitosis or worth a closer look
  Taken from 94 breast cancer slides
  Taken from 2,444 HPF

Question: Is this figure mitosis or not?
  Possible responses: Yes, No, Maybe
Agreement on mitotic figures

Pairwise agreement where both committed to “yes” or “no”:

Table 1: Pathologist agreement of mitotic figures.

<table>
<thead>
<tr>
<th>Pathologists</th>
<th>Yes/Yes</th>
<th>No/Yes</th>
<th>Yes/No</th>
<th>No/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/B</td>
<td>1352</td>
<td>4</td>
<td>789</td>
<td>102</td>
</tr>
<tr>
<td>A/C</td>
<td>2705</td>
<td>20</td>
<td>172</td>
<td>83</td>
</tr>
<tr>
<td>B/C</td>
<td>1506</td>
<td>756</td>
<td>15</td>
<td>461</td>
</tr>
</tbody>
</table>

Strong bias:

B seems to have much stricter criteria than C
C seems to have slightly stricter criteria than A
What about figure-level agreement?

Where do B and C agree? (randomly chosen examples)
Agreed “mitosis”
Where do B and C agree? (randomly chosen examples)

Agreed “not mitosis”
What about figure-level agreement?

Where do B and C disagree? (randomly chosen examples)

- B mitosis, C not mitosis (very few)
What about figure-level agreement?

Where do B and C disagree? (randomly chosen examples)

- **B not mitosis, C mitosis (many!)**
Hope: Train a machine that could behave as a consistent, reliable observer with good prognostic ability

Machine learning:

- Computer finds a rule (by solving a minimization problem) that allows it to predict pathologists' classification, in terms of values it can measure from the image pixels

- Which pathologist's classification to follow?
Candidate points and classification

SVR (support Vector Regression)
Map overall image staining level to color of mitotic figures

Color histogram

Color thresholding

Feature extraction

Feature vector

color
shape
mass
Inside texture

SVM+ (support Vector Machine)
Classify candidates as mitotic figures (red)

Trained model

SVR (support Vector Regression)
Map overall image staining level to color of mitotic figures

Trained model
Train the machine to predict the decision of the majority of the three pathologists

Exclude figures unless two pathologists agreed “yes” or two agreed “no”
- E.g. “yes”, “yes”, “no” => “yes”
- “yes”, “maybe”, “yes” => “yes”
- “yes”, “maybe”, “maybe” => exclude from data set

This is for investigation purposes, not a gold standard.
Divide data set into three parts (for training and model selection)

Results on test set (799 potential mitotic figures; at least two observers agreed):

<table>
<thead>
<tr>
<th>Observer</th>
<th>Majority label</th>
<th>Observer “Mitosis”</th>
<th>Observer “Maybe”</th>
<th>Observer “Not mitosis”</th>
<th>Lower bound agreement</th>
<th>Upper bound agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Positive (726)</td>
<td>658</td>
<td>65</td>
<td>3</td>
<td>90.6%</td>
<td>99.6%</td>
</tr>
<tr>
<td></td>
<td>Negative (73)</td>
<td>15</td>
<td>36</td>
<td>22</td>
<td>30.1%</td>
<td>79.4%</td>
</tr>
<tr>
<td>B</td>
<td>Positive (726)</td>
<td>394</td>
<td>166</td>
<td>166</td>
<td>54.3%</td>
<td>77.1%</td>
</tr>
<tr>
<td></td>
<td>Negative (73)</td>
<td>0</td>
<td>0</td>
<td>73</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>C</td>
<td>Positive (726)</td>
<td>720</td>
<td>4</td>
<td>2</td>
<td>99.2%</td>
<td>99.7%</td>
</tr>
<tr>
<td></td>
<td>Negative (73)</td>
<td>2</td>
<td>2</td>
<td>69</td>
<td>94.5%</td>
<td>97.3%</td>
</tr>
</tbody>
</table>

Recall that A and B disagreed on 793/2249 (35%) of the cases where they both committed to “yes” or “no”

In each of these cases, C’s vote determined the majority label

So it’s expected that he has high agreement
### How the machine predicts majority labels

**Results on test set:**

<table>
<thead>
<tr>
<th>Observer</th>
<th>Majority label</th>
<th>Observer “Mitosis”</th>
<th>Observer “Maybe”</th>
<th>Observer “Not mitosis”</th>
<th>Lower bound agreement</th>
<th>Upper bound agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Positive (726)</td>
<td>658</td>
<td>65</td>
<td>3</td>
<td>90.6%</td>
<td>99.6%</td>
</tr>
<tr>
<td></td>
<td>Negative (73)</td>
<td>15</td>
<td>36</td>
<td>22</td>
<td>30.1%</td>
<td>79.4%</td>
</tr>
<tr>
<td>B</td>
<td>Positive (726)</td>
<td>394</td>
<td>166</td>
<td>166</td>
<td>54.3%</td>
<td>77.1%</td>
</tr>
<tr>
<td></td>
<td>Negative (73)</td>
<td>0</td>
<td>166</td>
<td>73</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>C</td>
<td>Positive (726)</td>
<td>720</td>
<td>4</td>
<td>2</td>
<td>99.2%</td>
<td>99.7%</td>
</tr>
<tr>
<td></td>
<td>Negative (73)</td>
<td>2</td>
<td>2</td>
<td>69</td>
<td>94.5%</td>
<td>97.3%</td>
</tr>
<tr>
<td>Machine</td>
<td>Positive (726)</td>
<td>462</td>
<td>0</td>
<td>264</td>
<td>63.6%</td>
<td>63.6%</td>
</tr>
<tr>
<td></td>
<td>Negative (73)</td>
<td>1</td>
<td>0</td>
<td>72</td>
<td>98.6%</td>
<td>98.6%</td>
</tr>
</tbody>
</table>

Machine’s performance is similar to that of A and B
Even though A and B contributed to the “majority label”
How the machine predicts majority labels

True positives of the machine:
How the machine predicts majority labels

**True negatives** of the machine:
How the machine predicts majority labels

**False positives** of the machine:
False negatives of the machine:
**Conclusions**

- Evaluated pathologist agreement on *individual mitotic figures*, on an unprecedented scale

- Exposed systematic differences in the kinds of patterns counted

- Constructed an early-generation *computerized detector* that performs *within the bounds of real pathologists* in classifying mitotic figures

**Future work:**
- Special training for telophase
- Gold standard data set (compare with IHC staining)
- Calibrate to perform mitotic grading
- Investigate the prognostic ability of the computer’s output

- Supporting more efficient, more consistent grading
Empowered by Innovation

NEC