Формирование признаков распознавания гистологических изображений на основе стохастической геометрии и функционального анализа

Федотов Н. Г., Шульга Л. А., Кольчугин А. С., Смолькин О. А., Романов С. В.

ec@diamond.stup.ac.ru

Пенза, Пензенский государственный университет

Задача распознавания цитологических и гистологических изображений возникает при диагностике онкологических заболеваний. Суть цитологического и гистологического анализов заключается в подготовке препарата и рассмотрении его под микроскопом при различных увеличениях на предмет выявления морфологических признаков, характерных для онкологических заболеваний. В настоящей статье рассмотрена идея создания интеллектуальной системы диагностики, автоматически формирующей триплетные признаки распознавания. Данные признаки базируются на аппарате стохастической геометрии, эффективность которого была подтверждена в [1, 2]. Признаки распознавания в рассматриваемом подходе имеют структуру в виде композиции трех функционалов $\Pi(F) = \Theta \circ P \circ T\left(F \cap l(p,\theta)\right)$, где p, θ — нормальные координаты сканирующей прямой $l(p,\theta)$, с которыми связаны функционалы P и Θ; функционал T связан с естественной координатой t сканирующей прямой $l(p,\theta)$; и F — обозначение изображения распознаваемого объекта. В связи с характерной структурой такие признаки были названы триплетными, их подробное рассмотрение приведено в [2].

Применение данного аппарата непосредственно к исходным гистологическим изображениям затруднительно, поскольку на них изображены ядра, фолликулы, соединительная ткань и другие виды объектов, каждый из которых имеет свои собственные значимые характеристики. Триплетные признаки хорошо «схватывают» геометрические особенности изображенных объектов, но для этого сначала необходимо выполнить их выделение. Эта задача была решена предварительной обработкой изображений в соответствии с процедурой, описанной в [4]. В результате были получены отдельные изображения фолликул и отдельные изображения ядер препарата.

При практическом решении задачи распознавания всегда стоит проблема выделения наиболее информативных признаков. Триплетная структура позволяет получить тысячи различных признаков (для этого достаточно использовать всего 10 функционалов каждого типа), причем в режиме автоматической генерации. Однако вычислительная сложность получения такого числа признаков для каждого распознаваемого изображения, а также сложность построения решающей процедуры при таком

числе признаков требуют от нас ограничиться небольшим количеством наиболее информативных признаков.

Наш подход основывается на формальной генерации большого числа триплетных признаков, формируемых на основе имеющейся библиотеки функционалов, и последующем отборе согласно некоторому критерию эффективности как можно меньшего числа наиболее полезных для распознавания признаков. Отбор признаков часто называют процессом минимизации признакового пространства.

Минимизация обычно включает преобразование кластеризации и выбор признаков. Идея преобразования кластеризации заключается в том, чтобы обеспечить группировку точек, представляющих выборочные образы одного класса. В результате такого преобразования максимизируются расстояния между множествами и минимизируются внутримножественные расстояния.

С точки зрения теории информации критерием оптимизации выбора признаков может служить понятие энтропии. Признаки, уменьшающие неопределенность заданной ситуации, считаются более информативными, чем те, которые приводят к противоположному результату. Таким образом, если считать энтропию мерой неопределенности, то разумным правилом является выбор признаков, обеспечивающих минимизацию энтропии рассматриваемых классов. Это правило эквивалентно минимизации дисперсии в различных совокупностях образов, образующих классы. Выражения для энтропии дают полное представление об информативности описания. Но оценка по этим формулам затрудняется большим объемом вычислений, с учетом того, что в решаемой нами задаче изначально генерируется более 10000 признаков. Это делает задачу определения набора информативных признаков в рамках концепции минимизации энтропии неразрешимой за реальное время. Кроме того, концепция минимизации энтропии основывается на предположении о нормальности распределения образов, составляющих заданные классы, в то время как в реальных задачах законы распределений неизвестны. Объем обучающей выборки часто бывает небольшим, и делать оценки параметров распределения довольно рискованно. В этих условиях целесообразно использовать методы, которые не требуют построения модели распределения и опираются на объекты, имеющиеся в обучающей выборке.

Таким методом является разложение по системе ортогональных функций. При выборе признаков используют обобщенное разложение Карунена-Лоэва, поскольку оно обладает следующими оптимальными свойствами [3]:

1) минимизирует среднеквадратичную ошибку при использовании лишь конечного числа базисных функций в разложении;

2) минимизирует функцию энтропии, выраженную через дисперсии коэффициентов разложения.

Важность первого свойства заключается в том, что оно гарантирует невозможность получения меньшей в среднеквадратичном смысле ошибки аппроксимации с помощью другого разложения. Важность второго свойства заключается в том, что оно связывает с коэффициентами разложения оценку минимальной энтропии или дисперсии.

При генерации признаков распознавания для гистологических изображений изначально было получено 13 500 признаков. На предварительном этапе были отсеяны все вырожденные признаки, значения которых оказались постоянными для всех образов. К оставшимся признакам была применена процедура минимизации на основе разложения Карунена-Лоэва. В результате, для изображений фолликул при коэффициенте $k = 0.8$ было отобрано 59 признаков. Коэффициент k задает долю общей суммы дисперсий $D_j(E[f_{ji}])$ математических ожиданий всех признаков, которая обеспечивается за счет отобранных признаков. Соотношение внутриклассовых и межклассовых дисперсий для отобранных признаков позволяет эффективно организовать процедуру распознавания с использованием простых решающих правил.

Таким образом, можно сделать следующие выводы:

— применение признаков со структурой в виде композиции трех функционалов (триплетных признаков) позволяет формировать большое количество признаков в режиме автоматической компьютерной генерации;
— для отбора наиболее информативных признаков применима процедура, основанная на обобщенном разложении Карунена-Лоэва, которая обеспечивает минимизацию внутриклассовой энтропии, выражаемой через дисперсии коэффициентов разложения.

Работа выполнена при поддержке РФФИ, проект № 05-01-00991.

Литература

[1] *Федотов Н. Г.* Методы стохастической геометрии в распознавании образов. — М: Радио и связь, 1990.

[2] *Федотов Н. Г., Шульга Л. А.* Теория распознавания и понимания образов на основе стохастической геометрии // Искусственный интеллект. — 2002. — № 2. — С. 282–289.

[3] *Ту Дж., Гонсалес Р.* Принципы распознавания образов. — М.: Мир,1978.

[4] *Федотов Н. Г., Шульга Л. А., Кольчугин А. С., Романов С. В., Смолькин О. А., Курынов Д. В.* Предварительная обработка гистологических изображений в системе распознавания заболеваний щитовидной железы // сб. тр. «Надежность и качество–2006». — Пенза, 2006. — Т. 2. — С. 245–246.