



Здесь $\sum_{s \in S} \xi_s(t)$ - суммарное (с учетом знака) смещение всех соседних точек. При моделировании целесообразно принять $m=1$.

Литература

1. Басов К.А. ANSYS для конструкторов [Текст] / К.А. Басов. – М.: ДМК Пресс, 2009. – 248 с.
2. Дьяконов В.П. Matlab R2006/2007/2008. Simulink 5/6/7. Основы применения [Текст] / В.П. Дьяконов. – М.: Солон-Пресс, 2008. – 800 с.
3. Майер Р.В. Компьютерное моделирование физических явлений [Текст] / Р.В. Майер. – Глазов: ГГПИ, 2009. – 112 с.

Н.С. Кравцова, Р.А. Парингер, А.В. Куприянов

РАЗРАБОТКА МЕТОДОВ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ ДЕНДРИТНЫХ КРИСТАЛЛОГРАММ НА ОСНОВЕ ОЦЕНИВАНИЯ ФАКТОРОВ ФОРМЫ ПРОСТРАНСТВЕННОГО СПЕКТРА

(Самарский государственный аэрокосмический университет имени академика
С.П. Королева (национальный исследовательский университет))

Введение

Анализ изображений медицинских кристаллограмм – это важная часть медицинской диагностики [1]. Медицинские кристаллограммы – это структуры, образованные при кристаллизации солей вследствие высушивания биологической жидкости (слеза, кровь, слюна и т.д.). Автоматизация обработки изображений кристаллограмм позволит повысить качество диагностики и сократит затраты времени на её проведение. В данной работе используются алгоритмы расчёта признаков форм пространственного спектра. Для классификации изображений использовались два классификатора: классификатор по K-ближайшим соседям и Байесовский классификатор.

В данной работе применяются геометрические признаки формы. Расчёт данных признаков производится на основе геометрических характеристик объектов представленных на изображении.

Информация, содержащаяся в изображении кристаллограмм, является структурно избыточной [2]. Если рассматривать пространственный спектр таких изображений, то он будет расположен в узкой полосе частот, которую можно назвать несущей частотой. Так же известно, что если на исходном изображении преобладали параллельные полосы определённого направления, тогда и на Фурье-преобразовании будут исходного изображения будут преобладать полосы с таким же направлением. Это свойство можно использовать для анализа кристаллограмм (Рисунок 1).

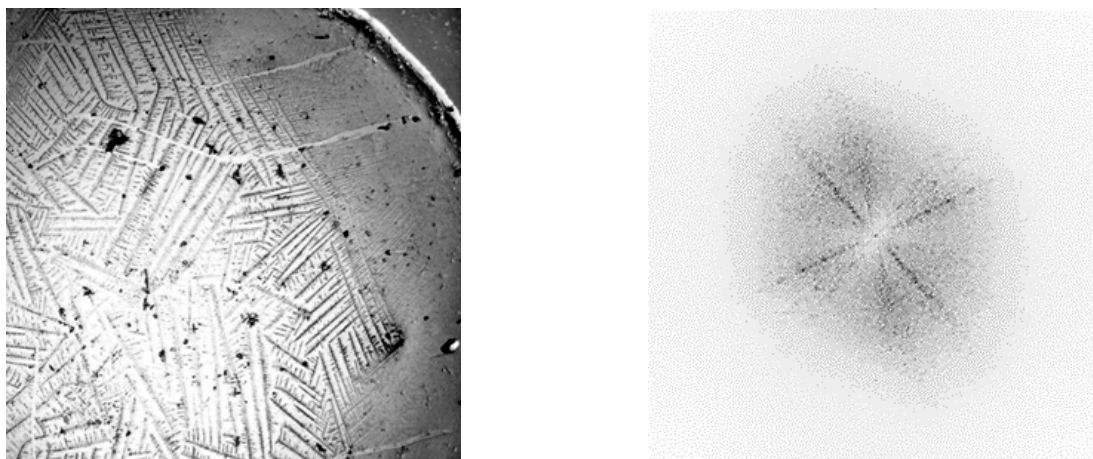


Рис. 1. Изображение кристаллограммы биологической жидкости и пространственного спектра (инвертированное изображение)

1 Факторы формы пространственного спектра

Информационная технология классификации изображений кристаллограмм с использованием факторов формы пространственного спектра заключается в следующем: выполняется выравнивание яркости исходного изображения, быстрое преобразование Фурье, построение гистограммы распределения яркости, пороговая обработка, преобразовании области в замкнутую, выделение границ, подсчёт факторов формы.

В данной работе рассчитываются различные геометрические признаки формы спектра.

1. Коэффициент компактности $F_1 = \frac{P^2}{S}$
2. Коэффициент округлости $F_2 = \sqrt{\frac{4S}{\pi}}$
3. Коэффициент формы $F_3 = \frac{4\pi S}{P^2}$
4. Коэффициент Малиновской $F_4 = \frac{P}{4\sqrt{\pi S}}$
5. Коэффициент Blair-Biss $F_5 = \frac{S}{\sqrt{2\pi \sum_i r_i^2}}$

где r_i – расстояние от точки до центра тяжести, P – периметр, а S – площадь.

Указанные признаки обладают свойством инвариантности т.е. неизменности величины по отношению к преобразованиям сдвиг, поворот, масштабирование.

Вычисление периметра P является нетривиальной задачей. В данной работе используется способ, состоящий из трёх этапов:

1) Применяется операция дилатации. Из полученного расширенного на один пиксель по всем направлениям изображения, попиксельно вычитаем исходное изображение. Получаем изображение «внешнего» контура.

2) Применяется операция эрозии к исходному. Из исходного изображения попиксельно вычитаем полученное изображение. Получаем изображение «внутреннего» контура.



3) Среднее значение двух полученных контуров можно использовать в качестве периметра Р.

Для подсчёта площади S применим более простой метод. Посчитаем количество белых точек на изображении. Если принять один пиксель за единицу площади. Тогда площадь будет равна сумме всех белых точек.

2 Байесовский классификатор.

Наивный байесовский классификатор – простой вероятностный классификатор, основанный на применении Теоремы Байеса со строгими (наивными) предположениями о независимости.

Достоинством наивного байесовского классификатора является малое количество данных для обучения, необходимых для оценки параметров, требуемых для классификации.

Пусть входной сигнал, задаваемый вектором $y = (y_0, \dots, y_{k-1})^T$ и подлежащий распознаванию, представляет собой аддитивную смесь детерминированной и шумовой составляющих. Будем считать, что наблюдаемые векторы имеют плотность вероятностей вида

$$p(y/\Omega_l) = \frac{1}{(2\pi)^{K/2} \sqrt{|B_l|}} \exp\left(-\frac{1}{2} (y-M_l)^T B_l^{-1} (y-M_l)\right), \quad l = \overline{0, L-1}$$

Здесь $M_l = E\{Y/\Omega_l\}$, $B_l = E\{(y-M_l)(y-M_l)^T/\Omega_l\}$ — математическое ожидание и ковариационная матрица вектора признаков из класса Ω_l , соответственно. Требуется по реализации y случайного вектора Y определить класс, к которому данный вектор принадлежит.

Решением данной задачи является классификатор Байеса с дискриминантными функциями следующего вида:

$$d_l(y) = \ln P(\Omega_l) - \ln \sqrt{|B_l|} - \frac{1}{2} (y-M_l)^T B_l^{-1} (y-M_l), \quad l = \overline{0, L-1}.$$

Данное выражение может быть упрощено. Предположим, что все корреляционные матрицы одинаковы. Тогда решающие функции классификатора Байеса представляются в виде

$$d_l(y) = 2 \ln P(\Omega_l) - (y-M_l)^T B_l^{-1} (y-M_l), \quad l = \overline{0, L-1}.$$

Величина $\rho(y, M_l) = (y-M_l)^T B_l^{-1} (y-M_l)$ называется расстоянием Махаланобиса между векторами y и M_l и является мерой близости вектора y к центру класса Ω_l .

3 Метод k ближайших соседей

Метод k ближайших соседей — метод автоматической классификации объектов. Основным принципом метода ближайших соседей является то, что объект присваивается тому классу, который является наиболее распространённым среди соседей данного элемента.

Алгоритм может быть применен к многомерным выборкам. Для этого перед применением нужно определить функцию дистанции. Классический вариант определения дистанции — дистанция в евклидовом пространстве.



4 Экспериментальные результаты классификации

Для обучения классификатора были использованы изображения формата jpg с разрешением 256 на 256 пикселей, состоящие из 256 полутоновых оттенков. Всего было выбрано 400 изображений, принадлежащих к разным классам. По 100 изображений каждого класса (Рисунок 2).



Рис. 2. Увеличенные фрагменты изображений дендритных кристаллограмм

Критерием качества классификации определим ошибку классификации, показывающую, в скольких случаях классификатор принял неправильные решения, по формуле: $\varepsilon = \frac{k}{n} \cdot 100\%$, где k – число ошибок классификации, n – общее число изображений.

Изображения исследуемых кристаллограмм были разбиты по группам: группа 1 – больные с поставленным диагнозом; группа 2 – после нескольких процедур лечения; группа 3 – группа прошедших курс лечения; группа 4 – группа «норма».

Результат эксперимента показал, что ошибка классификатора по k ближайшим соседям равна 11%. В случае использования Байесовского классификатора ошибка равна 16%.

Были построены детальные графики гистограмм распределения признаков по каждой группе кристаллограмм. Результаты для признака Малиновской представлены на рисунке 3. Пики на графиках определяют характерное значение признака для выбранной группы.

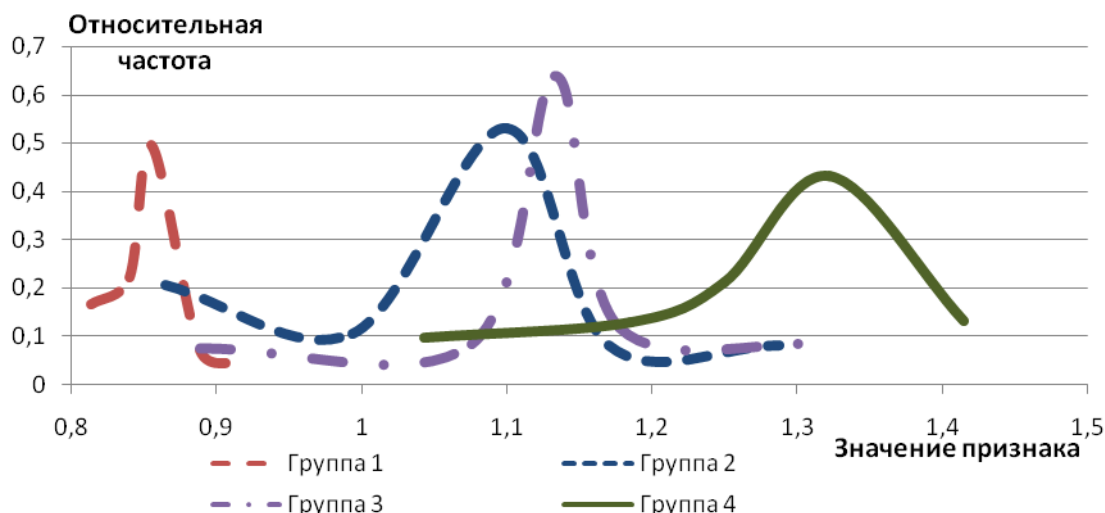


Рис. 3. Гистограмма распределения фактора Малиновской



Заключение

Анализ гистограмм распределения признаков пространственного спектра показал, что области распределения значений фактора Малиновской для различных групп кристаллограмм практически не пересекаются. Это свидетельствует об эффективности применения предложенной технологии анализа факторов формы пространственного спектра для классификации дендритных кристаллограмм.

Литература

1. Мартусевич, А. К. , Кристаллографический анализ: общая характеристика / А. К. Мартусевич // Вятский медицинский вестник. - 2002. - № 3. С. 59-61.
2. Куприянов, А. В. Сегментация текстурных изображений на основе оценивания локальных статистических признаков / А. В. Куприянов // Вестник СГАУ. – 2008. – №2 (15). – С. 245-252.

И.А. Лёзин, А.О. Авдиенко

РЕШЕНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ И МНОГОКРИТЕРИАЛЬНОГО ПОИСКА ПРИ ИМПОРТЕ БОЛЬШИХ МАССИВОВ ДАННЫХ

(Самарский государственный аэрокосмический университет имени академика С.П. Королева (национальный исследовательский университет))

По условию задачи, на вход автоматизированной системы поступает файл в некотором формате (например, CSV) с записями о сотрудниках организации, которые нужно импортировать в базу данных, предварительно провалидив и разделив на классы. Записи могут быть невалидными, повторяющимися, уже существующими в базе, новыми, и т.д. Объем файла довольно велик, несколько десятков тысяч строк.

Каждая запись содержит M полей, то есть в файле M колонок и N строк. Некоторые колонки используются для поиска дубликатов записей с учетом их приоритета.

Подробнее о возможных сложностях и проблемах:

- в файле могут быть записи вообще без колонок, по которым проводится проверка;
- имя задается в 3 колонках, при этом никто не может гарантировать, что фамилия будет стоять в графе "Фамилия";
- ни одна из поисковых колонок не уникальна среди всех записей, уникальность поддерживается только по группе атрибутов;
- файл импортируется не весь целиком, а по частям, поэтому непосредственно в момент валидации мы не видим всего файла целиком.

По результатам импорта формируется отчет.