Методы отбора признаков для задач классификации изображений земной поверхности

Е.Ф. Гончарова[а], А.В. Гайдель[а,б]

[а] *Самарский национальный исследовательский университет имени академика С.П. Королева, 443086, Московское шоссе, 34, Самара, Россия*
[б] *Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, 443001, ул. Молодогвардейская, 151, Самара, Россия*

Аннотация

В работе исследуются два метода отбора наиболее информативных признаков для повышения эффективности классификации изображений земной поверхности, полученных при дистанционном зондировании Земли. Один из рассмотренных методов отбора признаков основан на дискриминантном анализе, другой – на построении линейной регрессионной модели. В качестве характеристик изображений используются ряд гистограммных и текстурных признаков. Экспериментальное исследование предложенных методов на изображениях из базы данных UC-Merced Land Use показало, что с помощью отобранных признаков удается правильно классифицировать 95% изображений.

Ключевые слова: отбор признаков; классификация; геоинформатика; дискриминантный анализ; регрессионный анализ.

1. Введение

В настоящее время изображения, полученные с помощью дистанционного зондирования Земли (ДЗЗ), являются огромным источником данных. Исследование полученных изображений предоставляет возможность не только обогащать наши знания о планете, но и решать конкретные прикладные задачи. Например, контролировать обработку пахотных земель, следить за состоянием лесов и другие. Для решения всех этих задач необходимо развитие методов обработки изображений, предоставляющих возможность производить точную и эффективную классификацию изображений на классы.

Существует множество методов выделения и отбора наиболее информативных признаков, позволяющих получать хорошие результаты. Например, в работе [1] производилась классификация изображений на 19 классов на основании комбинации различных признаков. Средняя доля правильно классифицированных объектов составила 93,6%, для некоторых классов 100%. В работе [2] рассматривается вопрос сокращения признакового пространства в задачах распознавания на изображениях. Признаковое пространство, состоящее из нескольких сотен тысяч признаков (пикселов исходного изображения), было сокращено до нескольких десятков признаков.

Различные методы отбора наиболее информативных признаков широко используются при анализе биомедицинских изображений. Так, в работе [3] из 169 отдельных признаков, характеризующих протекание хронической обструктивной болезни ХОБЛ, с помощью метода, основанного на применении дискриминантного анализа, была выбрана группа из пяти признаков, которая позволила достичь вероятности ошибочной классификации в 0,11.

В данной работе предлагается исследовать гистограммные и текстурные признаки, характеризующие изображения. Изображения были получены из открытой базы данных UC Merced Land Use Dataset, которая предоставляет изображения размерностью 256×256 отсчетов, относящиеся к различным классам: поле, лес, пляж и другие. Отбор информативных признаков осуществляется с помощью двух методов: первый основан на дискриминантном, второй – на регрессионном анализе. Для проверки результатов, полученных с помощью двух предложенных методов, производится классификация методом ближайшего соседа.

2. Объект исследования

Объектом исследования в данной работе являются признаки, характеризующие изображения, а также методы, позволяющие производить отбор наиболее информативных признаков для последующей классификации изображений.

Предлагается исследовать гистограммные и текстурные характеристики изображения и степень влияния этих признаков на классификацию изображений на два класса.

С помощью методов отбора наиболее информативных признаков был выявлен набор характеристик, который необходимо учитывать для получения наилучших результатов классификации. Классификация проводится методом ближайшего соседа.

3. Методы

3.1. Формирование признаков

Изображение характеризуется своей матрицей яркости $I^{(M\times N)}$, где $M\times N$ – размер изображения в пикселях. Интенсивность каждого отсчета для изображения, представленного в формате RGB, определяется в соответствии со следующей формулой (1):

$$I(m,n)=\frac{R(m,n)+G(m,n)+B(m,n)}{3},\ \ m=\overline{1,M},\ \ n=\overline{1,N}, \tag{1}$$

где R,G,B – интенсивность красной, зеленой и синей составляющей отсчета с координатами (m,n) соответственно.

Существует множество признаков, которые являются характеристиками изображения. В данной работе были использованы гистограммные характеристики, являющиеся статистическими и описывающие распределение интенсивности. Если представить дискретное изображение как реализацию двумерного случайного процесса, то можно оценить распределение интенсивности изображения. Рассмотрим начальные моменты (2) и центральные моменты (3):

$$\nu_k=\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}I^k(i,j) \tag{2}$$

$$\mu_k=\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(I\left(i,j\right)-\nu_1\right)^k \tag{3}$$

В качестве характеристик изображения использовались:

- средняя интенсивность:
  $\overline{I}=\nu_1$, а также (I_R, I_G, I_B – средняя интенсивность по красным, зеленым и синим отсчетам соответственно);
- начальный момент второго порядка (средняя энергия):
  $s=\nu_2$;
- СКО (среднеквадратическое отклонение):
  $\sigma=\sqrt{\mu_2}$;
- коэффициент асимметрии:
  $\gamma_1=\frac{\mu_3}{\sigma^3}$;
- коэффициент эксцесса (мера остроты пика распределения случайной величины):
  $\gamma_2=\frac{\mu_4}{\sigma^4}-3$.

Рассмотрим автокорреляционную функцию (4), которая характеризует зависимость между отсчетами изображения.

$$R(m,n)=\frac{\frac{1}{(N-|m|)(M-|n|)}\sum_i\sum_j I(i,j)I(i+m,j+n)}{\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}I^2(i,j)}. \tag{4}$$

В качестве текстурных признаков использовались четыре отсчета (по четырем направлениям: горизонтальном, вертикальном и двух диагональных) автокорреляционной функции (4):

- $r_1=\frac{1}{2}\left(R\left(0,-1\right)+R\left(0,1\right)\right)$;
- $r_2=\frac{1}{2}\left(R\left(1,0\right)+R\left(-1,0\right)\right)$;
- $r_3=\frac{1}{2}\left(R\left(-1,-1\right)+R\left(1,1\right)\right)$;
- $r_4=\frac{1}{2}\left(R\left(1,-1\right)+R\left(-1,1\right)\right)$.

3.2. Методы отбора признаков

Пусть $\mathbf{\Omega}$ – множество объектов, подлежащих распознаванию. В данной задаче объектом этого множества является x_k – вектор признаков, характеризующий k -ое изображение. Множество было разбито на 2 класса с помощью разбиения $\mathbf{\Lambda} = \left\{ \mathbf{\Omega}_j \right\}_{j=1}^{2}$ так, что:

1) $\mathbf{\Omega}_0 \cup \mathbf{\Omega}_1 = \mathbf{\Omega}$,

2) $\mathbf{\Omega}_0 \cap \mathbf{\Omega}_1 = \varnothing$.

Пусть $\Phi(x_k) : \mathbf{\Omega} \to \mathbf{\Lambda}$ – идеальный оператор распознавания, который переводит объект распознавания x_k в его класс. Так как идеальный оператор нам неизвестен, необходимо построить другой оператор $\tilde{\Phi}(x_k) : \mathbf{\Omega} \to \mathbf{\Lambda}$, который, так же как и $\Phi(x_k)$ переводит объект распознавания в его класс, однако не владеет информацией о классе объекта. Для построения оператора $\tilde{\Phi}(x_k)$ необходимо воспользоваться информацией, получаемой из обучающей выборки $U \subseteq \mathbf{\Omega}$, для объектов которой класс является известным.

Перед проведением процедуры отбора необходимо провести стандартизацию векторов признаков для того, чтобы размерность каждого из признаков не влияла на результат отбора. Для этого необходимо оценить математическое ожидание:

$$M(i) = \frac{1}{|U|} \sum_{k=1}^{|U|} x_k(i), \quad i = \overline{1,L}, \quad M \in \mathbb{R}^L$$

где L – количество признаков,
и дисперсию:

$$R(i,i) = \frac{1}{|U|} \sum_{k=1}^{|U|} \left(x_k(i) - M(i) \right)^2, \quad i = \overline{1,L}, \quad R \in \mathbb{R}^{L \times L}$$

для каждого признака.

Затем после применения формулы (5) векторы признаков будут иметь нулевое математическое ожидание и единичную дисперсию.

$$x_k(i) = \frac{x_k(i) - M(i)}{\sqrt{R(i,i)}}, \quad k = \overline{1,|U|}, \quad i = \overline{1,L}. \tag{5}$$

Для отбора наиболее информативных признаков были рассмотрены два метода. Первый метод был предложен в работе [4] и основан на применении критерия дискриминантного анализа, согласно которому выбирается набор признаков, обеспечивающий максимум критерия $J(Q)$:

$$J\left(Q\right) = \frac{\operatorname{tr} R}{\sum_{j=1}^{2} \mathrm{P}\left(\Omega_j\right) \operatorname{tr} R_j},$$

где Q - текущий набор признаков;

R - корреляционная матрица смеси распределений;

R_j - корреляционная матрица внутри j -го класса;

$\mathrm{P}\left(\Omega_j\right)$ - вероятность попадания объекта из класса Ω_j, для данной задачи $\mathrm{P}\left(\Omega_j\right) = \frac{1}{2}$.

Таким образом, выбирается последовательность признаков, при которой рассеяние смеси распределений сильнее превышает среднее внутриклассовое рассеяние.

Второй метод отбора основан на применении регрессионного анализа. Регрессионный анализ изучает связь между зависимой переменной и одной или несколькими независимыми переменными.

В качестве зависимой переменной предлагается рассматривать номер класса $y\left(x\right)$, которому принадлежит тот или иной объект. Тогда $y\left(x\right)$ зависит от вектора признаков x. Было построено уравнение линейной регрессии:

$$y = X\theta + \xi,$$

где $y = (y_1 \quad y_2 \quad \ldots \quad y_n)^T$ – выходной вектор;

X - матрица признаков;

$\theta = \left(\theta_0 \quad \theta_1 \quad \ldots \quad \theta_{|Q|}\right)^T$ – вектор неизвестных коэффициентов регрессии;

$\xi = (\xi_1 \quad \xi_2 \quad \dots \quad \xi_n)^T$ – вектор ошибок, который содержит погрешности наблюдений.

Неизвестные коэффициенты, которые содержит вектор θ, были оценены по обучающей выборке методом наименьших квадратов:

$$(y - X\theta)^T (y - X\theta) \to \min_{\theta}.$$

Величина вклада каждого признака оценивается в соответствии с коэффициентом перед соответствующим признаком в уравнении регрессии.

Для проверки эффективности полученных наборов признаков была проведена классификация методом ближайшего соседа. На признаковом пространстве была определена евклидова метрика:

$$\rho(x, y) = \sqrt{\sum_{i=1}^{L} (x(i) - y(i))^2},$$

где L - количество признаков.

Классификатор определяет класс заданного вектора x, как класс его ближайшего соседа из обучающей выборки. Предложенный метод классификации достаточно прост с вычислительной точки зрения по сравнению с другими методами классификации. Для некоторых задач недостаток данного метода заключается в необходимости хранения большого количества данных обо всех объектах обучающей выборки и сравнении каждого из них с неизвестным объектом [5].

Для заданной системы распознавания вероятность ошибочного распознавания оценивается, как

$$\varepsilon = \frac{\left|\left\{ x_k \in \tilde{\mathbf{U}} \mid \Phi(x_k) \neq \tilde{\Phi}(x_k) \right\}\right|}{\left|\tilde{\mathbf{U}}\right|}, \; k = \overline{1, \left|\tilde{\mathbf{U}}\right|}.$$

где $\left|\tilde{\mathbf{U}}\right|$ – контрольная выборка.

4. Результаты

Для экспериментальной проверки предложенных методов были использованы два набора изображений из базы данных UC Merced Land Use Dataset, которая предоставляет изображения размерностью 256×256 отсчетов, относящиеся к различным классам: поле, лес, пляж и другие. В данной работе исследовались два класса изображений, характеризующие поле и лес. На рисунке 1 представлены примеры изображений из репозитория.

а) б)

Рис.1. Пример изображения из базы данных UC Merced Land Use (а - поле, б - лес).

В качестве обучающей выборки было использовано $\frac{4}{5}$ имеющихся изображений, для которых класс считался известным. Экспериментальная проверка проводилась на $\frac{1}{5}$ части изображений.

Результаты, полученные с помощью дискриминантного и регрессионного анализа, показаны в таблицах 1 и 2.

Таблица 1. Группы из первых 7 признаков, отобранных с помощью дискриминантного анализа, и вероятность ошибки

Признаки	ε
I_R	0,5
$I_R, \bar{I}$	0,075
$I_R, \bar{I}, s$	0,05
$I_R, \bar{I}, s, I_G$	0,075
$I_R, \bar{I}, s, I_G, I_B$	0,225
$I_R, \bar{I}, s, I_G, I_B, r_2$	0,175

$I_R, \bar{I}, s, I_G, I_B, r_2, r_1$	0,175

Таблица 2. Группы из первых 7 признаков, отобранных с помощью регрессионного анализа, и вероятность ошибки

Признаки	ε
I_R	0,5
I_R, I_G	0,075
$I_R, I_G, \bar{I}$	0,2
$I_R, I_G, \bar{I}, I_B$	0,175
$I_R, I_G, \bar{I}, I_B, r_1$	0,075
$I_R, I_G, \bar{I}, I_B, r_1, r_4$	0,1
$I_R, I_G, \bar{I}, I_B, r_1, r_4, s$	0,1

Таблица 3 представляет собой так называемую таблицу несоответствия для результатов классификации методом ближайшего соседа в пространстве, состоящем из трех признаков, отобранных с помощью метода дискриминантного анализа. Строки таблицы представляют собой реальные классы, к которым относятся объекты из обучающей выборки, столбцы таблицы указывают на предсказанные классификатором классы. По главной диагонали указана доля объектов, которые были правильно классифицированы для каждого из классов, и общая доля правильно классифицированных объектов для двух классов.

Таблица 3. Таблица классификации в пространстве из признаков $I_R, \bar{I}, s$

	поле	лес	
поле	100%	0%	
лес	10%	90%	
			95%

После рассмотрения результатов, полученных после оценки вероятности ошибочной классификации, можно прийти к выводу, что для предложенной задачи классификации изображений более эффективным оказался метод отбора признаков, основанный на дискриминантном анализе. Минимальная оценка вероятности ошибочной классификации 0,05 достигается на наборе из трех признаков $I_R, \bar{I}, s$. Исследованные текстурные признаки не оказали большое влияние на разделение изображений на два класса. Это связано с тем, что учитывалась взаимосвязь лишь двух соседних отсчетов изображения. Для того чтобы сделать вывод о влиянии данной группы признаков на разделимость классов, необходимо произвести расчет более сложных текстурных характеристик. В дальнейших работах будет исследовано большее количество текстурных признаков, а также разделение изображений на большее число классов, что, в свою очередь, может привести к изменению набора информативных признаков.

5. Заключение

Таким образом, для исходных данных, полученных для проведения исследования, были выявлены наиболее информативные признаки, которые необходимо учитывать при построении классификатора, способного предсказать класс, к которому относится изображение.

При анализе изображений, полученных в результате ДЗЗ, наиболее информативными оказались гистограммные признаки. Также стоит отметить, что данные изображения были представлены в формате RGB и включали в себя информацию о трех цветах. В связи с этим достаточно информативными оказались признаки, основанные на интенсивности по каждому из трех каналов.

Для набора изображений из базы данных UC-Merced Land Use лучший результат разделяющей способности (при проведении классификации методом ближайшего соседа) показали признаки, отобранные с помощью метода дискриминантного анализа. Минимальная оценка вероятности ошибочной классификации в этом случае оказалась равна 0,05, то есть доля верно классифицированных изображений составила 95%.

Метод дискриминантного анализа показал хорошие результаты и может применяться для отбора информативных признаков для задач классификации. В свою очередь стоит отметить, что было рассмотрено лишь небольшое количество признаков, характеризующих изображение, в дальнейших исследованиях будет рассмотрено большее количество различных признаков с разделением объектов на большее количество классов.

Благодарности

Работа выполнена при поддержке гранта РФФИ 16-41-630761 р_а, а также Министерства образования и науки РФ в рамках мероприятий Программы повышения конкурентоспособности Самарского университета среди ведущих мировых научно-образовательных центров на 2013-2020 годы и Программы фундаментальных исследований ОНИТ РАН «Биоинформатика, современные информационные технологии и математические методы в медицине».

Литература

[1] Guofeng Sheng. High-resolutionsatellite scene classification using a sparse coding based multiple featurecombination / Guofeng Sheng, Wen Yang, Tao Xu, Hong Sun // International Journal of Remote Sensing. – 2012. – vol. 33(8). – P. 2395-2412.

[2] Глумов, Н.И. Метод отбора информативных признаков на цифровых изображениях / Н.И. Глумов, Е.В. Мясников // Компьютерная оптика. – 2007. – Т. 31, № 3. – С. 73-76.

[3] Гайдель, А.В. Возможности текстурного анализа компьютерных томограмм в диагностике хронической обструктивной болезни / А.В. Гайдель, П.М. Зельтер, А.В. Капишников, А.Г. Храмов // Компьютерная оптика. – 2014. – Т. 38, № 4. – С. 843-850.

[4] Гончарова, Е.Ф. Статистическое исследование факторов, влияющих на развитие сердечно-сосудистых заболеваний / Е.Ф. Гончарова, А.В. Гайдель, А.Г. Храмов // сб. тр. конференции ИТНТ. – 2016. – С. 1020-1025.

[5] Fukunaga, K. Introduction to statistical pattern recognition / K. Fukunaga. – San Diego: Academic Press, 1990. – 592 p.