

ПРИМЕНЕНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ДЛЯ ЗАДАЧ ОБРАБОТКИ ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ЗЕМЛИ

Карпенко М.А.

Томский политехнический университет, Институт кибернетики
Научный руководитель: к.т.н. О.С. Токарева (ТПУ, Институт химии нефти СО РАН)
karpenkomaria8990@gmail.com

Введение

Обработка данных дистанционного зондирования Земли (ДЗЗ), представляемых в виде растровых изображений, включает в себя широкий ряд задач, одной из которых является классификация. Для классификации данных ДЗЗ чаще всего используются различные методы, использующие в качестве признаков спектральные характеристики пикселей в разных диапазонах электромагнитного спектра. Применение метода главных компонент (ГК) позволяет избежать дублирования информации, доставляемой взаимосвязанными признаками, малой вариативности признаков и предоставляет возможность агрегирования признаков [1].

Метод главных компонент

В исследовательской и практической работе приходится сталкиваться с ситуациями, когда общее число p признаков $x_1, x_2, \dots, x_p$, регистрируемых на каждом из множества обследуемых объектов, очень велико. Тем не менее, имеющиеся многомерные наблюдения следует подвергнуть статистической обработке. Представление каждого из наблюдений в виде вектора Z вспомогательных показателей $z_1, z_2, \dots, z_{p'}$ с существенно меньшим (чем p) числом компонент p' может быть обусловлено в первую очередь следующими причинами:

1. необходимостью визуализации исходных данных для более наглядного представления;
2. стремлением к лаконизму исследуемых моделей, обусловленному необходимостью упрощения счета и интерпретации полученных статистических выводов;
3. необходимостью существенного сжатия объемов хранимой статистической информации без видимых потерь в ее информативности.

При этом новые признаки могут выбираться из числа исходных или определяться по заданному правилу по совокупности исходных признаков [1].

Метод ГК – это один из способов понижения размерности, состоящий в переходе к новому ортогональному базису, оси которого ориентированы по направлениям максимальной дисперсии набора входных данных. Вдоль первой оси нового базиса дисперсия максимальна, вторая ось максимизирует дисперсию при условии ортогональности первой оси, и т.д., последняя ось имеет минимальную дисперсию из всех возможных. Такое преобразование позволяет понижать информацию путем отбрасывания

координат, соответствующих направлениям с минимальной дисперсией [2].

Геометрически определение первой ГК равносильно построению новой координатной оси OZ_1 таким образом, чтобы она шла в направлении наибольшего разброса данных, т.е. в направлении вытянутости анализируемого «облака» многомерных наблюдений. Затем, среди направлений, перпендикулярных к OZ_1 , отыскивается направление «наибольшей вытянутости» OZ_2 и т.д. Направление этих новых осей представляет собой направление собственных векторов матрицы ковариации исходных признаков.

Таким образом, формально, преобразование по методу ГК можно представить следующим образом:

$$Z = P^T * X,$$

где X – матрица исходных признаков, размерности $p \times n$, p – количество исходных признаков, n – количество обследуемых объектов;

P^T – транспонированная матрица преобразований размерности $p' \times p$, p' – количество ГК;

Z – матрица ГК, размерности $p' \times n$.

Матрица преобразований P представляет собой матрицу размерности $p \times p'$, где в столбцы записаны собственные вектора матрицы ковариации исходных данных, записанные по порядку уменьшения соответствующих собственных значений [3].

Алгоритм работы программы

С использованием метода ГК реализована программа, в основе которой лежит следующий алгоритм:

1. На первом этапе производится расчет вектора средних значений. Эти данные необходимы для расчета матрицы ковариаций;
2. Рассчитывается матрица ковариаций исходной матрицы;
3. Выполняется расчет собственных векторов и соответствующих собственных значений векторов;
4. Формируется матрица преобразований, для этого собственные вектора упорядочиваются по их собственным значениям, и выбираются m собственных векторов с наибольшими собственными значениями, где m – количество запрошенных ГК на выходе;
5. Исходная матрица признаков преобразуется в матрицу ГК с помощью матрицы преобразования, полученной на предыдущем этапе.

Первая главная компонента (ГК1) будет иметь максимальное количество независимой информации, второй главная компонента (ГК2) будет иметь второе по величине количество независимой информации, не содержащейся в первой главной компоненте и т.д. [4].

Разработанная программа протестирована на наборе данных «Ирисы Фишера». Ирисы Фишера состоят из данных о 150 экземплярах ириса, по 50 экземпляров на каждый из трёх видов. Также было проведено сравнение результатов обработки изображений с результатами, получаемыми при использовании программного комплекса ERDAS Imagine.

На выходе программа визуализирует распределение набора ирисов в двумерном пространстве первой и второй ГК.

Результаты

Метод ГК применен для обработки космического снимка (КС) со спутника Landsat-7 с пространственным разрешением 30 м. На рис. 2 приведены изображения данных КС, полученные в 1-5 и 7 каналах.

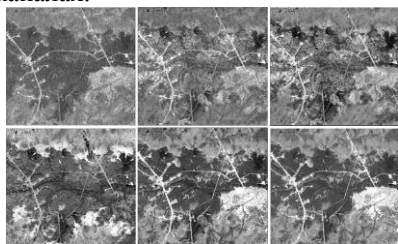


Рис. 2. Изображения данных КС в 1-5 и 7 каналах. На рис. 3 приведены изображения ГК, полученные в результате обработки данных КС по методу ГК.

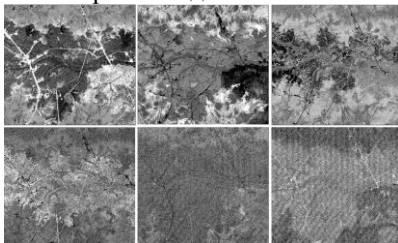


Рис. 3. Компоненты 1-6 КС после обработки

В результате разложения на ГК можно увидеть, что наиболее вариативная информация содержится в ГК1-3, менее вариативная – в компонентах 4-6.

Далее проведена классификация данных КС на основе спектральных признаков пикселей в исходных каналах (рис. 4) и на основе признаков в пространстве первых трех ГК, полученных после обработки КС по методу КГ (рис. 5). Разными цветами на рис. 4 и 5 обозначены следующие объекты: синий, зеленый – лес разного типа, черный – антропогенные объекты, белый – гари, розовый, голубой – болотная растительность.

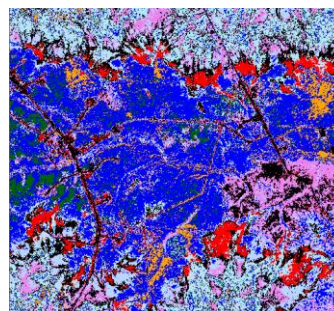


Рис. 4. Результаты классификации данных КС на основе значений спектральной яркости пикселей

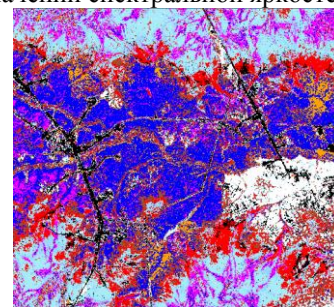


Рис. 5. Результаты классификация КС на основе значений первых трех главных компонент

Сравнение результатов классификации КС и карты данной территории показало, что в результате классификации с использованием в качестве признаков значений ГК полученные кластеры более правильно описывают имеющиеся на данной территории объекты. Как видно на рис. 5 более отчетливо выделены антропогенные объекты, зоны гарей и болот.

Заключение

В дальнейшем планируется использование разработанной программы для снижения размерности признакового пространства при использовании текстурных признаков изображения и значений спектральных яркостей пикселей, получаемых при съемке территории в разное время вегетационного периода растительного покрова.

Список использованной литературы

1. Айвазян С. А. и др. Прикладная статистика. Классификации и снижение размерности. – Финансы и статистика, 1989.
2. Померанцев А. Метод главных компонент (PCA) //Российское хеометрическое общество.[Электронный ресурс].–Режим доступа: <http://rco.chemometrics.ru/Tutorials/pca.htm> (дата обращения: 23.12. 2013).
3. Kegl B. Principal manifolds for data visualization and dimension reduction. – Berlin-Heidelberg : Springer, 2008. – Т. 1.
4. Полищук Ю. М., Перемитина Т. О. Геоинформационный комплекс анализа состояния окружающей среды на основе метода главных компонент //Вычислительные технологии. – 2004. – Т. 9.– С. 14-25.