УДК 004.93’1; 004.932

Д.В. Прадун

СНИЖЕНИЕ РАЗМЕРНОСТИ ОБУЧАЮЩИХ ВЫБОРОК ПРИ РАСПОЗНАВАНИИ ОБРАЗОВ НА КОСМИЧЕСКИХ ИЗОБРАЖЕНИЯХ С ПОМОЩЬЮ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

Описываются сущность метода главных компонент и задача снижения размерности в процессе статистической обработки. Приводится способ вычисления главных компонент на основе определения собственных значений ковариационной матрицы. Описываются алгоритмы практической реализации метода главных компонент на основе QR-алгоритма. Проводится анализ возможности использования метода главных компонент при классификации космических изображений с целью снижения размерности обучающих выборок.

Введение

Методы классификации космических изображений с использованием обучающих выборок наряду с классификацией по эталону представляют собой наиболее распространенные способы группировки исследуемых объектов в классы. Процедура обучения играет важную роль при реализации алгоритма распознавания образов, так как от качества ее выполнения зависит корректность отнесения образов к тому или иному классу. Для этого обучающая выборка должна содержать достаточную для классификации информацию о классах и эталонах, к которым те или иные анализируемые образы относятся. Данное требование не всегда может быть выполнено, так как имеющиеся базы обучающих выборок часто содержат «избыточную» информацию о признаках анализируемых объектов, которая при использовании в процессе обучения ухудшает качество классификации и увеличивает время обработки.

Метод главных компонент (МГК) как способ снижения размерности многомерных данных при классификации статистических данных применяется в мировой практике довольно широко на протяжении нескольких лет. Примерами использования данного метода являются разработки по распознаванию лиц на изображениях [1–3], исследования в области спектроскопии [4], охраны окружающей среды [5] и многих других. Привлекательность МГК в тех или иных областях исследований обусловлена относительной простотой его использования, а также хорошей теоретической и практической проработкой при реализации метода в соответствующих системах анализа статистических данных.

В статье приведено общее техническое решение возможности использования МГК как процедуры снижения размерности (объема) обучающих выборок при классификации с обучением. Данное исследование было сделано в первую очередь для проверки влияния МГК на качество и быстродействие реализации классификаторов с обучением, а также возможности параллельной модификации метода.

1. Сущность метода главных компонент

МГК представляет собой один из способов снижения размерности анализируемых многомерных признаков, широко используемый в статистике [6, 7]. Наряду с факторным анализом и многомерным шкалированием, представляющими собой альтернативные методы снижения размерности, МГК нашел широкое применение именно при решении прикладных статистических задач [8], а также реализован в некоторых программных пакетах. Сущность МГК состоит в следующем.

Пусть имеются многомерные статистические данные вида

$$X_i = \begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \dots \\ x_i^{(p)} \end{pmatrix}, \; i = 1, 2, ..., n \, . \tag{1}$$

Если число p слишком велико, возникает задача представления каждого наблюдения из (1) в виде вектора Z некоторых вспомогательных показателей $z^{(1)}, z^{(2)}, ..., z^{(p')}$ со значением $p' << p$. Другими словами, при определенной p'-мерной вектор-функции $Z = Z(X)$ исходных переменных $x^{(1)}, x^{(2)}, ..., x^{(p)}$ задается мера информативности $I_{p'}(Z(X))$ p'-мерной системы признаков $Z = \left(z^{(1)}(X), z^{(2)}(X), ..., z^{(p')}(X) \right)$ [6, 7].

Задача МГК, как и других методов снижения размерности, заключается в определении такого набора признаков $\widetilde{Z}$, найденного в классе $\boldsymbol{F}$ допустимых преобразований исходных показателей $x^{(1)}, x^{(2)}, ..., x^{(p)}$, что

$$I_{p'}(\tilde{Z}(X)) = \max_{Z \in F} \{ I_{p'}(Z(X)) \} \, . \tag{2}$$

Для этого определим в качестве класса $F(X)$ все возможные линейные ортогональные нормированные комбинации, такие, что

$$\begin{aligned} & z^{(j)}(X) = c_{j1}(x^{(1)} - \mu^{(1)}) + ... + c_{jp}(x^{(p)} - \mu^{(p)}); \\ & \sum_{v=1}^{p} c_{jv}^2 = 1, \; j = 1, 2, ..., \; p; \\ & \sum_{v=1}^{p} c_{jv} c_{kv} = 0, \; j, k = 1, 2, ..., \; p, \; j \neq k, \end{aligned} \tag{3}$$

где $\mu^{(v)}$ – математическое ожидание $x^{(v)}$; c_{jv} – произвольные постоянные числа [6]. Мера информативности p'-мерной системы признаков $\left(z^{(1)}(X), z^{(2)}(X), ..., z^{(p')}(X) \right)$ в этом случае определяется как

$$I_{p'}(Z(X)) = \frac{\boldsymbol{D} z^{(1)} + ... + \boldsymbol{D} z^{(p')}}{\boldsymbol{D} x^{(1)} + ... + \boldsymbol{D} x^{(p)}} \, , \tag{4}$$

где $\boldsymbol{D}$ – дисперсия соответствующей случайной величины [6, 7].

При таком определении класса преобразований $F(X)$ для любых фиксированных $p' = 1, 2, ..., p$ вектор вспомогательных переменных $\tilde{Z}(X) = \left(\tilde{z}^{(1)}(X), \tilde{z}^{(2)}(X), ..., \tilde{z}^{(p')}(X) \right)$ определяется как линейная комбинация $\widetilde{Z} = \boldsymbol{L} X$, где

$$\boldsymbol{L} = \begin{pmatrix} l_{11} \dots l_{1p} \\ \dots\dots\dots\dots \\ l_{p'1} \dots l_{p'p} \end{pmatrix}, \tag{5}$$

а ее строки удовлетворяют условию ортогональности. При этом

$$I_{p'}(\tilde{z}^{(1)}(X), ..., \tilde{z}^{(p')}(X)) = \max_{Z(X)\in F} I_{p'}(Z(X)).$$

Полученные таким образом переменные $\tilde{z}^{(1)}(X), \tilde{z}^{(2)}(X), ..., \tilde{z}^{(p')}(X)$ являются главными компонентами вектора X. Таким образом, *первой главной компонентой* $\tilde{z}^{(1)}(X)$ исследуемой системы показателей $X = (x^{(1)}, x^{(2)}, ..., x^{(p)})'$ называется такая нормированно-центрированная линейная комбинация этих показателей, которая среди всех прочих нормированно-центрированных линейных комбинаций переменных $x^{(1)}, x^{(2)}, ..., x^{(p)}$ обладает наибольшей дисперсией. Тогда *k-я главная компонента* $(k = 2, 3, ..., p)$ исследуемой системы показателей $X = (x^{(1)}, x^{(2)}, ..., x^{(p)})'$ представляет собой такую нормированно-центрированную линейную комбинацию этих показателей, которая не коррелирована с предыдущими $(k-1)$ главными компонентами и среди всех прочих нормированно-центрированных и не коррелированных с предыдущими $(k-1)$ главными компонентами линейных комбинаций переменных $x^{(1)}, x^{(2)}, ..., x^{(p)}$ обладает наибольшей дисперсией [6, 7].

Поскольку решение задачи, т. е. вид матрицы линейного преобразования $\boldsymbol{L}$, зависит только от элементов ковариационной матрицы $\boldsymbol{\Sigma}$, которые, в свою очередь, не изменяются при замене исходных переменных $x^{(j)}$ переменными $x^{(j)} - c^{(j)}$, исходную систему показателей необходимо центрировать, т. е. $Ex^{(j)} = 0, j = 1, 2, ..., p$. В статистической практике этого добиваются переходя к наблюдениям $\tilde{x}_i^{(j)} = x_i^{(j)} - \bar{x}^{(j)}$, где $\bar{x}^{(j)} = \sum_{i=1}^{n} x_i^{(j)} / n$.

Из определения главных компонент следует, что для вычисления первой главной компоненты необходимо решить оптимизационную задачу вида

$$\begin{cases} \boldsymbol{D}(l_1 X) \to \max\limits_{l_1}; \\ l_1 l_1' = 1, \end{cases} \tag{6}$$

где l_1 – первая строка матрицы L из формулы (5). Учитывая центрированность переменной X ($\boldsymbol{E}X = 0$) и то, что $\boldsymbol{E}(XX') = \boldsymbol{\Sigma}$, имеем $\boldsymbol{D}(l_1 X) = \boldsymbol{E}(l_1 X)^2 = \boldsymbol{E}(l_1 X X' l_1') = l_1 \boldsymbol{\Sigma} l_1'$.

Следовательно, задача (6) может быть записана следующим образом:

$$\begin{cases} l_1 \boldsymbol{\Sigma} l_1' \to \max\limits_{l_1}; \\ l_1 l_1' = 1. \end{cases} \tag{7}$$

Решая первое уравнение системы через функцию Лагранжа, получаем систему уравнений для определения l_1:

$$(\boldsymbol{\Sigma} - \lambda \boldsymbol{I}) l_1' = 0. \tag{8}$$

Для того чтобы существовало ненулевое решение системы (8) (а оно должно быть ненулевым, так как $l_1 l_1' = 1$), матрица $\Sigma - \lambda \boldsymbol{I}$ должна быть вырожденной, т. е.

$$\left| \boldsymbol{\Sigma} - \lambda \boldsymbol{I} \right| = 0. \tag{9}$$

Этого добиваются за счет подбора соответствующего значения λ. Уравнение (9) относительно λ называется характеристическим для матрицы $\boldsymbol{\Sigma}$. Известно, что при симметричности

и неотрицательной определенности матрицы $\mathbf{\Sigma}$, каковой она и является как всякая ковариационная матрица, это уравнение имеет p вещественных неотрицательных корней $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, называемых характеристическими (собственными) значениями матрицы $\mathbf{\Sigma}$.

Учитывая, что $\boldsymbol{D}z^{(1)} = \boldsymbol{D}(l_1 X) = l_1 \sum l_1'$ и $l_1 \mathbf{\Sigma} l_1' = \lambda$ (последнее следует из (8) после умножения слева на l_1 с учетом $l_1 l_1' = 1$), получаем

$$\boldsymbol{D}z^{(1)}(X) = \lambda .$$

Подставляем λ_1 в систему уравнений (8) и, решая ее относительно $l_{11}, ..., l_{1p}$, определяем компоненты вектора l_1. В этом случае первая главная компонента представляет собой линейную комбинацию $z^{(1)}(X) = l_1 X$, где l_1 – собственный вектор матрицы $\mathbf{\Sigma}$, соответствующий наибольшему собственному значению этой матрицы [6, 7].

Далее аналогично можно показать, что $z^{(k)}(X) = l_k X$, где l_k – собственный вектор матрицы Σ, соответствующий *k*-му по величине собственному значению l_k этой матрицы.

Таким образом, соотношения для определения всех p главных компонент вектора X могут быть представлены в виде $Z = LX$, где $Z = (z^{(1)}, ..., z^{(p)})'$, $X = (x^{(1)}, ..., x^{(p)})'$, а матрица $\boldsymbol{L}$ состоит из строк $l_j = (l_{j1}, ..., l_{jp}), \quad j = \overline{1, p}$, являющихся собственными векторами матрицы $\mathbf{\Sigma}$, соответствующими собственным числам λ_j. При этом матрица $\boldsymbol{L}$ по построению является ортогональной, т. е. $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{L}'\boldsymbol{L} = 1$ [6, 7].

2. Практическая реализация метода главных компонент

Исходя из описанного выше вычисления главных компонент можно сделать вывод, что практическая реализация МГК сводится к задаче вычисления собственных значений ковариационной матрицы $\mathbf{\Sigma}$. Решения подобной задачи стали предлагать еще в 1960-х гг., и они основывались на использовании классического QR-алгоритма [9]. Данный алгоритм позволял вычислять собственные значения заданной матрицы A не слишком высокого порядка n, не принадлежащей ни к какому специальному классу матриц и хранящейся в виде квадратного массива данных размером $n \times n$ в оперативной памяти [9].

Под QR-алгоритмом можно понимать итерационный процесс вида

$$A_k = Q_k R_k, \; R_k Q_k = A_{k+1}, \tag{10}$$

где $k = 1, 2, ...$, R_k – правая (верхняя) треугольная матрица на *k*-й итерации при условии, что столбцы матрицы A_k берутся в порядке $a_1, a_2, ..., a_n$. Если же столбцы берутся в обратном порядке, можно говорить о так называемом QL-алгоритме [10], т. е.

$$A_k = Q_k L_k, \; L_k Q_k = A_{k+1}.$$

Главным недостатком QR- и QL-алгоритмов была высокая трудоемкость. Чтобы ее существенно снизить, использовали следующие решения [9]:

1) приведение матрицы A к хессеберговой форме, которая сохраняется на протяжении всех QR-итераций;

2) использование сдвигов, т. е. выполнение QR-итераций для матриц вида $A_k - \mu_k I$ при подходящим образом выбранных числах μ_k;

3) проведение двойных QR-итераций в случае, если матрица A является вещественной.

Перечисленные приемы были реализованы в QR-алгоритме, представленном в программном пакете EISPACK [6, 7].

Как уже было сказано, QR-алгоритм предназначен для вычисления собственных значений обычных квадратных матриц. Если же матрица A является разреженной, т. е. имеет большое количество нулевых элементов, то, как показали исследования [9], QR-алгоритм с какого-то момента не сможет хранить промежуточные результаты применения итераций. Поэтому были предложены альтернативы QR-алгоритму, такие как методы одновременных итераций, метод Ланцоша, метод Арнольди и др. Так, метод Ланцоша был реализован во многих програм-мных пакетах, однако версии его программной реализации часто имели жесткие ограничения к применению и не получили дальнейшего развития [9].

С развитием вычислительных средств и появлением многопроцессорных высокопроизводительных систем появились новые возможности в реализации QR-алгоритма. В первую очередь это связано с его параллельной реализацией, которая была предложена в работах [11, 12]. Распараллеливание QR-итераций позволяет существенно ускорить процесс вычисления собственных значений матриц в случае их большого размера.

3. Использование метода главных компонент при классификации космических изображений

Как уже упоминалось, МГК является одним из эффективных способов значительного сжатия исходной информации при ее статистической обработке [6]. Применительно к процессу распознавания образов на изображениях это означает, что МГК способен сократить размеры обучающих выборок при использовании классификации с обучением с целью обеспечения классификаторов информацией, достаточной для обучения и не являющейся избыточной, а также сокращения времени ее обработки.

Для проверки данного утверждения использовался простейший классификатор на основе байесовских дискриминантных функций [13]. Применительно к классификатору была определена база обучающих выборок для класса зеленых насаждений. Кроме того, были заданы обучающие выборки для класса объектов, которые формально можно описать как «не принадлежащие зеленым насаждениям». Входные данные для МГК представляют собой некоторое число наборов критериев для классификации набора пикселов космических снимков в некоторой системе категорий («зеленые насаждения»). Число наборов критериев равняется количеству снимков, для которых предварительно были сформированы наборы пикселов, удовлетворяющих критериям. Таким образом, входные данные представляют собой несколько пар обучающих выборок (каждая пара – для отдельного снимка), содержащих набор пикселов снимка. В каждой паре один файл обучающей выборки содержит пикселы снимка, которые принадлежат зеленым насаждениям, а второй файл содержит пикселы снимка, которые не принадлежат зеленым насаждениям (рис. 1).

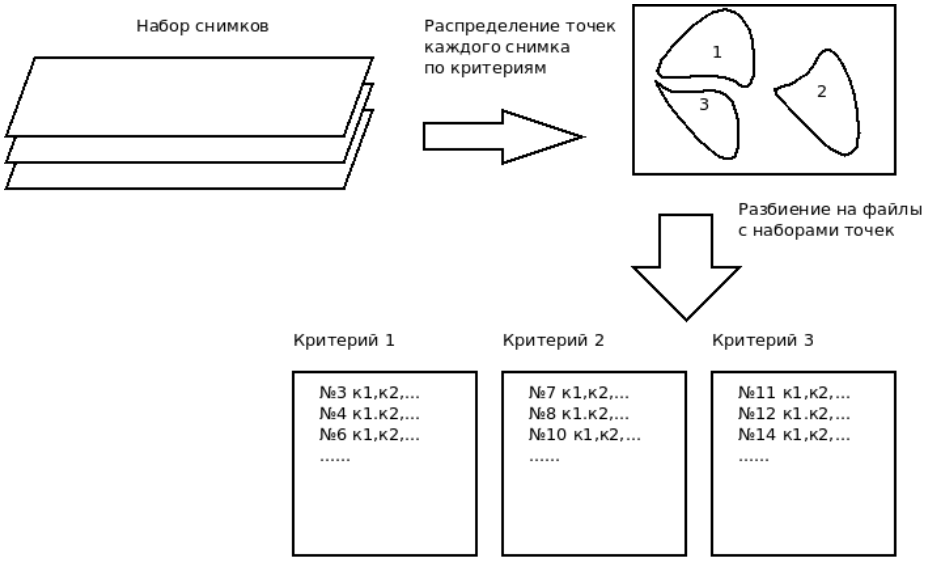


Рис. 1. Структура входных данных

Формат каждого файла обучающей выборки:

$$< N > : < K1 >, \ldots, < Kn >,$$

где N – индекс пиксела; $K1, ..., Kn$ – интенсивность пиксела по каждому каналу исходного изображения.

Для программной реализации МГК использовалась программная библиотека LAPACK, которая является продолжением пакета EISPACK [14, 15]. При этом реализация МГК выполнялась в так называемом псевдопараллельном режиме: исходные данные обучающих выборок делились на блоки фиксированного размера, для каждого блока применялся МГК для расчета собственных векторов матрицы блока и затем происходило формирование усредненной матрицы по всем блокам для каждого цветового канала и критерия. Процесс разбиения на блоки и формирования усредненной матрицы использовался, во-первых, во избежание возможного переполнения оперативной памяти при обработке обучающих выборок больших размеров, а во-вторых, для проверки возможности распараллеливания программной реализации МГК.

Байесовский классификатор применялся для каждого цветового слоя космического изображения отдельно. При этом предварительно выполнялась фильтрация исходного снимка с помощью параллельного алгоритма максимального потока сети [16], а затем кластеризация отфильтрованного изображения параллельным алгоритмом нечеткой кластеризации [17]. Данная предобработка осуществлялась с целью облегчения процесса распознавания образов с точки зрения количества пикселов, исходные характеристики которых подавались на вход байесовского классификатора. Другими словами, за счет группировки пикселов в кластеры существенно снизилось количество анализируемых данных во время процесса классификации. После обработки каждого цветового слоя на выходе получается массив флагов принадлежности V каждого пиксела изображения заданным классам объектов. Если выполняется условие

$$V_{i,j} > \frac{L}{2},\ i = 1, ..., N,\ j = 1, ..., L, \tag{11}$$

где L – количество цветовых слоев снимка, а N – максимальный индекс пиксела в изображении, то пиксел с индексом i определяется как принадлежащий зеленым насаждениям.

Аналогично байесовскому классификатору метод главных компонент применялся для каждого отдельного цветового слоя анализируемого снимка. В этом случае средний объем данных, подаваемый на вход МГК по обучающим выборкам каждого класса, равен K / L, где K – общий размер базы обучающих выборок для каждого класса.

В результате проведенных тестов (табл. 1) было установлено, что использование МГК позволяет значительно сократить объемы оперативной памяти, необходимой для обучения классификатора и дальнейшего распознавания образов с его помощью. Вместе с тем за счет выполнения дополнительной операции в виде процедуры снижения размерности обучающих выборок общее время обработки исходного снимка увеличивается. Поэтому выбор того, использовать МГК в процессе классификации с обучением при наличии большой базы обучающих выборок или нет, зависит от технических характеристик вычислительной системы, на которой выполняется процесс обработки. Тем не менее, если время обработки не является критическим параметром, использование МГК вполне оправдано.

Таблица 1

Показатели работы последовательного и блочно-параллельного алгоритмов

Средний объем данных обучающих выборок для цветового слоя, Мб	Время обработки*		Объем оперативной памяти, Мб	
	МГК и Байес	Байес	МГК и Байес	Байес
51,0	1 мин 16 с	45 с	24,3	35,0
518,8	7 мин 51 с	5 мин 54 с	76,4	345,6

* На вычислительной платформе Intel Core 2 Quad 2,66 ГГц, 3,24 Гб ОЗУ

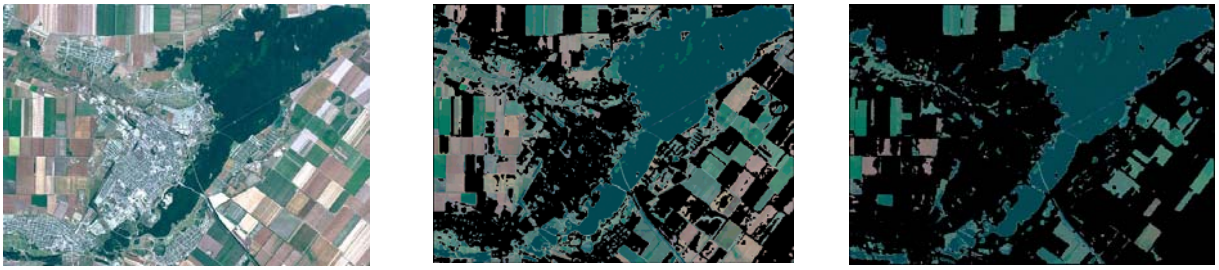
В отличие от байесовского классификатора, для которого дополнительно выполнялась кластеризация данных на основе нечеткой логики, при использовании МГК в процессе классификации с помощью метода *K* ближайших соседей [18] время обработки исходных данных существенно сокращается (табл. 2).

Таблица 2

Показатели работы метода *K* ближайших соседей и МГК

Время обработки	
МГК и *K* ближайших соседей	*K* ближайших соседей
22 с 985 мс	1 мин 41 с 828 мс
7 мин 4 с	15 мин 49 с

При этом качество классификации может существенно повыситься (рис. 2). Отсюда можно сделать вывод, что эффективность использования МГК в процессе классификации с обучением повышается при выполнении классификации исходных, не обработанных предварительно графических данных. В других же случаях использование МГК зависит от технических характеристик вычислительной системы, на которой выполняется процесс обработки. При этом если время обработки не является критическим параметром, использование метода главных компонент вполне оправдано.



а) *б)* *в)*

Рис. 2. Результаты классификации методом *K* ближайших соседей (*K* = 8): *а)* исходный снимок; *б)* классический метод; *в)* использование МГК для снижения размерности обучающей выборки

Заключение

МГК как один из способов снижения размерности многомерных данных позволяет существенно сократить размер базы обучающих выборок при реализации процесса классификации космических изображений. Его использование приводит к увеличению общего времени обработки данных, однако в случае когда временной фактор не является критическим, данным показателем работы МГК можно пренебречь. Кроме того, за счет разбиения входных данных обучающих выборок на блоки для каждого отдельного цветового слоя мультиспектрального изображения появляется возможность параллельной реализации МГК. При этом МГК может не только сокращать объем обучающих выборок, но и в отдельных случаях улучшать качество классификации.

Список литературы

1. Thakur, S. Face recognition using Principal Component Analysis and RBF Neural Networks / S. Thakur [et al.] // IJSSST. – 2009. – Vol. 10, № 5. – P. 7–15.

2. Zhang, D. Diagonal Principal Component Analysis for Face Recognition / D. Zhang, Z.-H. Zhou, S. Chen // Pattern Recognition. – 2006. – Vol. 39, № 1. – P. 140–142.

3. Bidyanta, N. Pattern Recognition using Principal Component Analysis / N. Bidyanta // Binary Digits [Electronic resource]. – 2010. – Mode of access : https://sites.google.com/site/binarydigits10/articles/eigenface. – Date of access : 10.09.2012.

4. Xiaoli, L. A Novel Approach to Pattern Recognition Based on PCA-ANN in Spectroscopy / L. Xiaoli, H. Yong // Lecture Notes in Computer Science. – 2006. – Vol. 4093. – P. 525–532.

5. Ferraz, A. The use of principal component analysis (PCA) for pattern recognition in Eucalyptus grandis wood biodegradation experiments / A. Ferraz [et al.] // World Journal of Microbiology and Biotechnology. – 1998. – Vol. 14, № 4. – P. 487–490.

6. Айвазян, С.А. Прикладная статистика. Классификация и снижение размерности. Справочное издание / С.А. Айвазян [и др.]. – М. : Финансы и статистика, 1989. – 608 с.

7. Айвазян, С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. – М. : ЮНИТИ, 1998. – 1005 с.

8. Дронов, С.В. Многомерный статистический анализ : учебное пособие / С.В. Дронов. – Барнаул : Изд-во Алт. гос. ун-та, 2003. – 213 с.

9. Икрамов, Х.Д. Несимметричная проблема собственных значений. Численные методы / Х.Д. Икрамов. – М. : Наука, 1991. – 240 с.

10. Парлетт, Б. Симметричная проблема собственных значений. Численные методы; пер. с англ. / Б. Парлетт. – М. : Мир, 1983. – 384 с.

11. Stewart, G.W. A parallel implementation of the QR algorithm / G.W. Stewart // CiteSeerX - Scientific Literature Digital Library and Search Engine [Electronic resource]. – University park, USA, 1987. – Mode of access : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.4400. – Date of access : 22.08.2012.

12. Henry, G. A parallel implementation of the nonsymmetric QR algorithm for distributed memory architectures / G. Henry, D. Watkins, J. Dongarra // SIAM Journal on Scientific Computing. – 2002. – Vol. 24, № 1. – P. 284–311.

13. Гонсалес, Р. Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – М. : Техносфера, 2005. – 1072 с.

14. Smith, B.T. Matrix eigensystem routines – EISPACK guide / B.T. Smith [et al.] // Lecture Notes in Computer Science. – 1976. – Vol. 6. – 551 p.

15. Garbow, B.S. Matrix eigensystem routines – EISPACK guide extension / B.S. Garbow [et al.] // Lecture Notes in Computer Science. – 1977. – Vol. 51. – 343 p.

16. Прадун, Д.В. Блочно-параллельная кластеризация мультиспектральных изображений с помощью алгоритма максимального потока в сети / Д.В. Прадун, Б.А. Залесский // Информатика. – 2011. – № 2(30). – С. 12–20.

17. Прадун, Д.В. Блочно-параллельная кластеризация изображений на основе нечеткой логики / Д.В. Прадун, А.А. Кравцов // Пятый Белорусский космический конгресс : материалы конгресса. В 2 т. (25-27 октября 2011 года, Минск). – Минск : ОИПИ НАН Беларуси, 2011. – Т. 2. – С. 47–53.

18. Hastie, T. The Elements of Statistical Learning. Data mining, Inference, and Prediction (Second Ed.) / T. Hastie, R. Tibshirani, J. Friedman // Trevor Hastie – Publications [Electronic resource]. – 2009. – Mode of access : http://www.stanford.edu/~hastie/local.ftp/ Springer/ESLII_print5.pdf. – Date of access : 12.03.2012.

Поступила 23.08.12

Объединенный институт проблем
информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail:pradundv@gmail.com

D.V. Pradun

REDUCTION OF TRAINING SAMPLES DIMENSION IN PATTERN RECOGNITION OF SPACE IMAGES USING PRINCIPAL COMPONENTS ANALYSIS

The essence of principal components analysis and the problem of dimension reduction are described. A method of principal components calculation is presented, which is based on the covariance matrix eigenvalues determination. Practical implementations of principal components analysis are described, which are based on QR-algorithm. Application of principal components analysis in space images classification for the reduction of training samples dimension is discussed.