УДК 004.934.2

МЕТОД ГЛАВНЫХ КОМПОНЕНТ И ЛИНЕЙНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ ДЛЯ СНИЖЕНИЯ РАЗМЕРНОСТИ В СИСТЕМАХ РАСПОЗНАВАНИЯ ОБРАЗОВ

В.В. МИШУГИНА

Белорусский государственный университет информатики и радиоэлектроники
ул. П. Бровки, 6, г. Минск, 220013, Республика Беларусь
mishugina@bsuir.by

При построении систем распознавания образов одним из важнейших этапов является выбор системы признаков. С помощью оптимального выбора из первоначального набора признаков можно удалить несущественные и избыточные данные. Для решения задачи выбора существенных признаков применяются различные методы и алгоритмы. В работе проводится сравнение метода главных компонент (Principal Component Analysis, PCA) и линейного дискриминантного анализ (LDA – Linear Discriminant Analysis).

Ключевые слова: системы распознавания образов, уменьшение размерности, метод главных компонент, линейный дискриминантный анализ.

Задачи, возникающие при построении автоматической системы распознавания образов, можно обычно отнести к нескольким основным областям. Одна из них связана с выделением характерных признаков или свойств из полученных исходных данных и снижением размерности векторов образов. Эффективность классификации образов зависит от информативности выбранного множества признаков. От мощности этого множества, т. е. размерности пространства признаков, существенно зависит скорость распознавания и объем необходимой информации. Понизить размерность пространства признаков можно за счет удаления малоинформативных признаков, но без существенного снижения вероятности правильного распознавания. Понижение размерности часто происходит неявно во всех модулях системы распознавания: предварительной обработки, выделения признаков и классификации.

Метод главных компонент [1] и линейный дискриминантный анализ пространства признаков [2], основанные на вычислении собственных чисел ковариационных матриц и применении преобразования Карунена–Лоэва, позволяют понизить размерность пространства признаков и улучшить кластеризацию образов.

Метод главных компонент рассматривает как одно целое всю совокупность образов, принадлежащих к разным классам. Вначале по всем имеющимся образам, относящимся к разным классам, вычисляется усредненный вектор признаков $x_{cp} = \frac{1}{P}\sum_{j=1}^{P} x_j$, где P – полное число образов; x_j – вектор признаков j-го образа, и формируется центрированная матрица исходных образов $D_{NP} = \left[(x_1 - x_{cp})^T, (x_1 - x_{cp})^T, \ldots, (x_1 - x_{cp})^T\right]$, где N – число признаков, которое и надо сократить.

Далее для D_{NP} вычисляется ковариационная матрица cov_{NN}, что в векторной форме можно представить как $\mathrm{cov}_{NN} = D_{NP} D_{NP}^T$. Для матрицы cov_{NN} определяются собственные числа и соответствующие им собственные векторы, для которых выполняется условие $\Lambda_{NN} = V_{NN}^T \mathrm{cov}_{NN} V_{NN}$, где Λ_{NN} – диагональная матрица, на диагоналях которой находятся собственные числа матрицы cov_{NN}; V_{NN} – ортогональная матрица, строки которой определяют собственные векторы, соответствующие собственным числам, т.е. $V_{NN} V_{NN}^T = I$, где I – единичная матрица.

Из собственных чисел, стоящих на диагонали матрицы Λ_{NN}, необходимо отобрать *k* наибольших, после чего в матрице V_{NN}^{T} оставить *k* строк, соответствующих этим числам, получив матрицу преобразования Карунена-Лоэва F_{kN}. Применив данное преобразование к каждому из *P* образов, можно получить значения данного образа в сокращенном пространстве признаков, т. е. $y_f^T - F_{kN} x_f^T$. В результате размерность пространства признаков сокращается.

Метод линейного дискриминантного анализа также используется для сокращения числа признаков, он заключается в следующем. Вначале вычисляются средний вектор признаков для всех образов x_{cp} средние векторы признаков для каждого *l*-го класса x_{cp}^{l} и формируются центрированные матрицы $D_{NT_l}^{l}$ образов для каждого класса аналогично методу главных компонент. В общем случае число образов *l*-го класса T_l у каждого класса разное, следовательно, число столбцов в матрицах D может быть различным.

Далее вычисляются средняя внутриклассовая ковариационная матрица $\mathrm{cov} = \sum_{l=1}^{M} D_{NT_l}^{l} (D_{NT_l}^{l})^{T}$ и межклассовая ковариационная матрица $\mathrm{cov}_{KI} = \sum_{l=1}^{M} (x_{cp}^{i} - x_{cp})(x_{cp}^{i} - x_{cp})^{T}$, причем размерность этих матриц совпадает и равна числу признаков N. На основе этих матриц вычисляется обобщенная ковариационная матрица $H_C = \mathrm{cov}^{-1}\mathrm{cov}_{KI}$, для которой определяются собственные числа и соответствующие им собственные векторы, причем, $\Lambda = V^{T} H_C V$, где Λ – диагональная матрица собственных чисел; V– ортогональная матрица, строки которой определяют собственные векторы, соответствующие собственным числам.

В диагональных элементах матрицы Λ необходимо выбрать *s* наибольших собственных чисел и преобразовать матрицу V^{T}, оставив в ней только соответствующие *s* строк. Полученную в результате матрицуA_{sN} можно использовать для преобразования всех образов $y^{T} = A_{sN} x^{T}$. В результате размерность пространства признаков сокращается.

Несмотря на то, что метод главных компонент и метод дискриминантного анализа повсеместно используются для понижения размерности пространства признаков, оба метода имеют свои преимущества и недостатки. Метод главных компонент относительно легко реализовать, так как матрица, используемая в разложении по собственным векторам всегда невырожденная. Что не относится к методу дискриминантного анализа. Кроме того, метод главных компонент требует меньше вычислений, особенно когда вычисляется собственное пространство для каждого класса и формируется основанный на МГК классификатор без использования других методов классификации образов. Однако дискриминантная информация не может постоянно находиться в направлении максимального значения дисперсии. Вот где слабое место метода главных компонент и где блистает метод дискриминантного анализа. ЛДА страдает от проблемы сингулярности, когда база данных обучения, является небольшой. Тем не менее, так как методы могут взаимно дополнять друг друга, наиболее успешным решением, является сочетание линейного дискриминантного анализа с методом главных компонент.

Список литературы

1. *Pearson K.,* On lines and planes of closest fit to systems of points in space // Philosophical Magazine, 1901, P. 559–572;

2. *Fisher, R. A.,* The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics, 1936, Vol 7 (2), P. 179–188.

3. *Кухарев Г. А.,* Биометрические системы: Методы и средства идентификации личности человека. СПб, 2001.

4. *Ерош И. Л., Сергеев М. Б., Соловьев Н. В.,* Обработка и распознавание изображений в системах превентивной безопасности: учеб. пособие, СПб, 2005.

5. *Martinez, A. and Kak, A.,* PCA versus LDA // Pattern Analysis and Machine Intelligence, 2001, IEEE Transactions on 23(2), P. 228–233.