УДК 519.6

ПРИМЕНЕНИЕ ПОДХОДОВ, СВЯЗАННЫХ С МИНИМИЗАЦИЕЙ ДИСПЕРСИИ, В ЗАДАЧАХ КЛАССИФИКАЦИИ МНОГОМЕРНЫХ ДАННЫХ БИОМЕДИЦИНСКОЙ ПРИРОДЫ

Туровский Я.А., *к.м.н., зав. лабораторией медицинской кибернетики Воронежского государственного университета, e-mail:*_____;
Борзунов С.В., *к.ф.-м.н., доцент кафедры цифровых технологий Воронежского государственного университета, e-mail:*_____;
Белобродский В.А., *аспирант кафедры цифровых технологий Воронежского государственного университета, e-mail:* _____.

DISPERSION MINIMIZATION APPROACHES IN BIOMEDICAL MULTIVARIABLE DATA QUALIFICATION

Turovskyi Ya.A., Borzunov S.V., Belobrodskyi V.A.

The focus of this article is an innovative method of digital data processing based on factor analysis, developed and tested to solve the task of data categorization using customized principal component analysis (PCA). The method searches the n-space for a new coordinate system, where component variance of a class is minimal. The result is that a data cluster is formed, in which probability density of a given class data is significantly higher than the one of the data belonging to other classes.
The accuracy of the method was verified by sequential enumeration of all possible angles of a data cloud rotation in a given coordinate system.

Key words: digital processing of signals, customized principal component analysis, biomedical signal classification.

Разработан и апробирован метод цифровой обработки данных на основе факторного анализа, обеспечивающий решение задачи классификации данных на основе адаптированного метода главных компонент. Суть метода заключается в поиске в n-мерном пространстве новой системы координат, дисперсия проекций данных одного из классов на ось которой будет минимальна. При этом формируется кластер, плотность вероятности нахождения наблюдений данного класса в котором существенно выше плотности вероятности нахождения в этом кластере данных из других классов. Произведена верификация метода при помощи альтернативного решения задачи методом последовательного перебора всех возможных углов поворота облака данных в существующей системе координат.

Ключевые слова: цифровая обработка сигналов, адаптированный метод главных компонент, классификация биомедицинских сигналов.

Введение

Анализ значительного числа разнородных данных является одной из традиционных задач большинства научных исследований в самых разнообразных сферах деятельности [1-3]. Для её решения разработан широчайший арсенал методов, доказавших свою эффективность [4-10]. Тем не менее, совершенствование уже существующих и поиск новых методов является одним из приоритетных направлений в области обработки результатов исследований, включая сигналы различной природы. Традиционные подходы классификации данных основаны на разделении наблюдений, относящихся к разным классам, т.е. выделении границ (возможно с использованием нечёткой логики), разделяющих в пространстве области, относящиеся к разным классам, и, одновременно, объединении наблюдений, расположенных «близко» в заданной, по сути, созданной под задачи исследования, метрике.

Целью работы является создание и апробация метода цифровой обработки данных на основе подходов минимизации дисперсии класса при использовании метода главных компонент.

Пусть исследуемая выборка представляет собой конечное множество объектов $\{\boldsymbol{x}_i\}$, $i = 1,2,...,N$, каждый из которых представляется вектором вещественнозначных признаков $\boldsymbol{x}_i$. Как хорошо известно, см., например, [11], в случае успеха процедура классификации разбивает выборку на непересекающиеся группы (кластеры):

$$\{\boldsymbol{x}_j\} = \bigcup_{i=1}^{K} C_i,\ \forall i \neq (C_i \cap C_j = \varnothing) \qquad (1)$$

При этом, анализируется вся исследуемая выборка, часть из которой используется для обучения классификатора, а часть – для кросс-проверки, обеспечивая избегание переобученности. Таким образом, по сути, можно сформулировать, что один из подходов к классификации сводится к тому, чтобы найти область (области) в уже существующем пространстве, или получить новые координаты в пространстве, в которых вероятность нахождения объектов одного класса будет статистически значимо больше, чем в других областях.

В отличие от широко распространённого подхода с разделением выборки, содержащей все исследуемые классы, на «обучающую» и «кросс-проверочную» под-

выборки, в рассматриваемом подходе будет использован иной алгоритм. Пусть из множества классов в массиве наблюдений имеется один, выделение которого значимо для целей и задач исследования. Или же имеется несколько групп, значимость нахождения конкретного наблюдения в каждой из которых, не равнозначна исходя из целей и задачей исследования. Рассмотрим единственный кластер C_1, который в дальнейшем применим для обучения классификатора. Следовательно, задача сводится к преобразованию существующих координат наблюдений в новые, обеспечивающие выполнение условий минимизации дисперсии точек кластера C_1. Преобразуем координаты существующего пространства так, чтобы в новых координатах проекция интересующего нас класса на одну из новых осей была минимальной. Т.е. рассмотрим другой подход, задачей которого является максимизация числа наблюдений одного из классов на единицу новой оси пространства наблюдений (в общем случае гиперплоскости). Иными словами, необходимо так изменить исследуемые координаты переменных, чтобы проекции всех векторов одного класса на новую ось координат (гиперплоскость) имели бы минимальный вариационный размах:

$$\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i,\boldsymbol{w}_k)^2 \to \min \ \text{ для } k=1,2,...n \qquad (2)$$

где $\boldsymbol{x}_i$ – центрированные векторы данных ($i = 1,\ldots,N$), $\boldsymbol{w}_k$ – векторы, характеризующие искомые оси новой системы координат, n – размерность пространства векторов исходных данных.

Для верификации статистической значимости различия величин дисперсии будем использовать критерий Ливена, который, по сравнению с критерием Бартлетта, менее чувствителен к отличиям распределения выборки от нормального распределения. Согласно критерию Ливена для выборки $\{\boldsymbol{x}_i\}$, $i = 1,2,\ldots,N$, разделенной на классы $\{\boldsymbol{x}^{(j)}_i\}$, $i = 1,2,\ldots,k$ вычислим величину

$$W = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k}N_j(\overline{Z}_{i\bullet}-\overline{Z}_{\bullet\bullet})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_j}(\overline{Z}_{ij}-\overline{Z}_{i\bullet})^2}, \qquad (3)$$

где $Z_{ij} = \left|(\boldsymbol{x}_i^{(j)},\boldsymbol{w}_1) - M_i\left[(\boldsymbol{x}_i^{(j)},\boldsymbol{w}_1)\right]\right|$, $M_i[...]$, – среднее по элементам i-го класса, $\overline{Z}_{i\bullet}$ – групповое среднее Z_{ij}, $\overline{Z}_{\bullet\bullet}$ – среднее по всей выборке. Гипотеза о равенстве дисперсий отвергается в случае $W > F_{\alpha,k-1,N-k}$, где через $F_{\alpha,k-1,N-k}$ обозначена верхняя граница F-распределения с k-1 и N-k степенями свободы на уровне значимости α.

При этом изменение координат векторов других классов ($C_{2..n}$) должно обеспечивать более низкую плотность вероятности их нахождения в области вариационного размаха переменной выбранной для обучения.

Таким образом, необходимо выполнение условия:

$$\frac{\sum_{i=1}^{N_j}(\boldsymbol{x}_i^{C_j},\boldsymbol{w}_k)^2}{\left|\max(\boldsymbol{x}_i^{C_j},\boldsymbol{w}_k)-\min(\boldsymbol{x}_i^{C_j},\boldsymbol{w}_k)\right|} < A(C_j), \qquad (4)$$

где $A(C_j) = \text{const}$.

Обозначим левую часть формулы (4) через $P(C_j)$ – это плотность точек класса C_j. Таким образом, вычисление этой величины методом градиентного спуска можно записать как:

$$\boldsymbol{x}_i^{C_j,(k)} = \boldsymbol{x}_i^{C_j,(k-1)} - \lambda^{(k-1)}\nabla P(C_j), \qquad (5)$$

где λ определяет скорость градиентного спуска.

Однако, возможно и применение иных методов для поиска новых координат. Как известно, примером преобразования, в результате которого могут быть изменены координаты в N-мерном пространстве тех или иных наблюдений, полученных экспериментально, может служить факторный анализ. Этот широко применяемый мощный метод анализа основан на поиске, т.н. латентных переменных – новых осей в уже существующем пространстве переменных, вдоль которых дисперсия облака данных максимальна [12, 13].

Следовательно, используя аппарат факторного анализа, можно получить данные о проекции облака данных на рассчитанные оси, при этом в ряде случаев дисперсия этих проекций будет последовательно увеличиваться:

$$0 < \lambda(w_1) < \lambda(w_2) < \ldots < \lambda(w_n). \qquad (6)$$

Очевидно, что новые оси (факторы), являющиеся собственными векторами ковариационной матрицы исходных переменных формируют множество двухмерных плоскостей. При этом вектор ортогональный вектору, дисперсия проекции переменных на который максимальна, имеет, наоборот, минимальную проекцию в данной плоскости (см. (2)).

Не составляет труда найти из множества собственных векторов (факторов, латентных переменных) тот, дисперсия проекции на который является минимальной. При этом расчеты следует вести не по ковариационной матрице всех наблюдений, а только по ковариационной матрице того класса, который выбран для формирования новых координат. Пусть $\boldsymbol{x}_i$ – записанные по столбцам, образуют матрицу $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$, и векторы, используемые для построения новых осей координат, имеют индексы $j = p, p+1,...,N$, где $1 < p < N$. Ковариационная матрица $\tilde{\Sigma}$ системы векторов $\tilde{X} = (x_1,...,x_{p-1})$ строится как

$$\tilde{\Sigma} = M[(x_l, x_m)] \text{ для } l,m \in [1, p-1], \qquad (7)$$

где M – математическое ожидание. Полученные собственные векторы этой матрицы формируют матрицу преобразования A, которая отражает преобразование исходного пространства в новое, так, что синхронно меняются и координаты всех переменных:

$$x' = A \cdot x. \qquad (8)$$

Лемма. Для точек кластера C_1 в декартовом пространстве R^2 перейдем к новой системе координат, оси которой Ox' и Oy' направлены вдоль собственных векторов $\boldsymbol{w}_1$ и $\boldsymbol{w}_2$ соответственно, причем этим векторам отвечают собственные значения $0 < \lambda(\boldsymbol{w}_1) < \lambda(\boldsymbol{w}_2)$. Тогда среднеквадратичное отклонение координат $\boldsymbol{x}'$ принимает минимальное значение среди всех возможных систем координат.

Доказательство леммы следует из того факта, что в методе главных компонент при выборе собственных векторов в соответствии с убыванием отвечающих им собственных значений происходит максимизация суммы квадратов проекций на первую ось («главная компонента»). Выбор обратного порядка собственных значений приводит к минимизации дисперсии.

Обобщение леммы на случай произвольного арифметического векторного пространства R^N не составляет трудности: выбирается вектор $\lambda(\boldsymbol{w}_1)$ с наименьшей (но, разумеется, большим нулю) собственным значением, причем остальные векторы из набора $\widetilde{X}$ располагаются в произвольном порядке, например, в порядке возрастания $\lambda(\boldsymbol{w}_i)$. Сформулированный алгоритм приводит к минимальному значению среднеквадратичного отклонения координат $\boldsymbol{x}'$ для требуемого класса данных.

Верификация подхода

Далее рассмотрим на практике поиск осей с минимальной дисперсией для объекта, представляющего собой три пересекающихся облака точек (см. рис. 1).

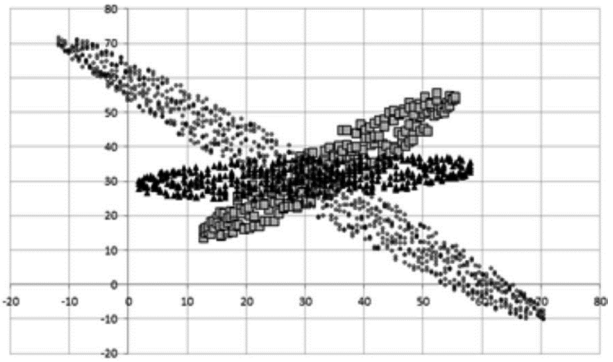

Рис. 1. Три облака точек на двумерной плоскости

Для верификации представленного алгоритма произведем два независимых друг от друга численных эксперимента (реализующие разные подходы) и сравним полученные результаты. В первом эксперименте применим подход, основанный на представленном выше алгоритме, то есть на формировании поворотной матрицы, составленной из собственных векторов ковариационной матрицы $\tilde{\Sigma}$ в порядке возрастания модуля соответствующих собственных чисел. Во втором численном эксперименте возьмем те же самые входные данные (см. рис. 1) и реализуем следующий итерационный алгоритм:

– на каждой итерации цикла от 0 до 360^0 с шагом в 1 градус будем производить:

• поворот 2-х мерного облака на текущий угол (от 0 до 360^0) с помощью известной матрицы поворота (матрицы направляющих косинусов) [14]:

$$A(0)=\begin{pmatrix}\cos\theta & \mp\sin\theta\\ \pm\sin\theta & \cos\theta\end{pmatrix}$$

• вычисление дисперсии первой компоненты $\boldsymbol{x}_1$ всего облака и запись пары (угол-дисперсия) в отдельный массив;

– после окончания итерационного процесса перебора всех возможных углов проанализируем полученные пары (угол-дисперсия) и найдем среди них угол, соответствующий минимальной дисперсии.

Исходя из взаимного расположения «облаков» (см. рис. 2 *а, б*) и соответствующих им числовых значений дисперсий можно сделать вывод о правильности работы алгоритма и его реализации.

При этом, как нетрудно заметить, дисперсия распределения остальных данных («облаков») в результате проекции на новую ось оказывается существенно выше, чем для выбранного класса.

Заключение

Разработан новый метод классификации данных, основанный на адаптации метода главных компонент для задачи выделения ограниченного числа кластеров. Данный метод позволяет, за счёт изменения системы координат, выделить оси, проекции на которые дают минимальную дисперсию исследуемого класса. При этом, в случае успешного обучения, дисперсии остальных классов по критерию Ливена оказываются значимо выше, чем дисперсия требуемого класса. Отсюда следует, что плотность вероятности проекции данных исследуемого класса на новую ось оказывается статистически значимо выше, чем для других классов выборки. Данный метод был протестирован и с помощью альтернативного решения, основанного на итерационном переборе, и показано достижение сходных результатов.

Работа выполнена при *поддержке РФФИ (грант № 19-07-01037 А).*

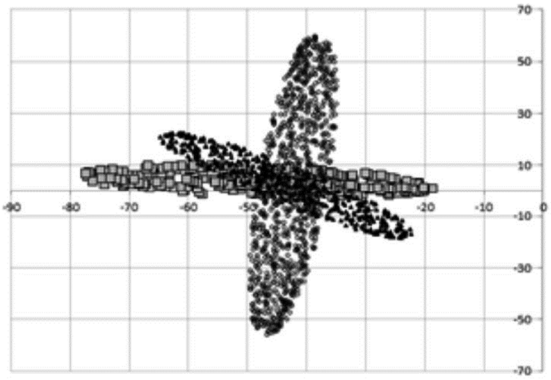

Рис. 2 а. Расположение «облаков», найденное с помощью представленного метода ($\sigma_X^2=92$, $\sigma_Y^2=648$)

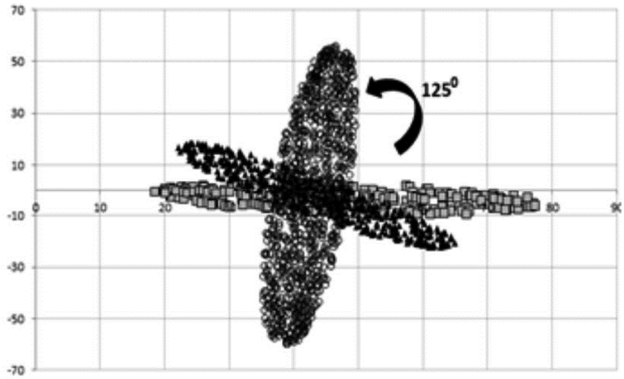

Рис. 2 б. Расположение «облаков», найденное с помощью последовательного перебора всех возможных углов ($\sigma_X^2=92$, $\sigma_Y^2=648$, $\alpha_{\min}=125^0$)

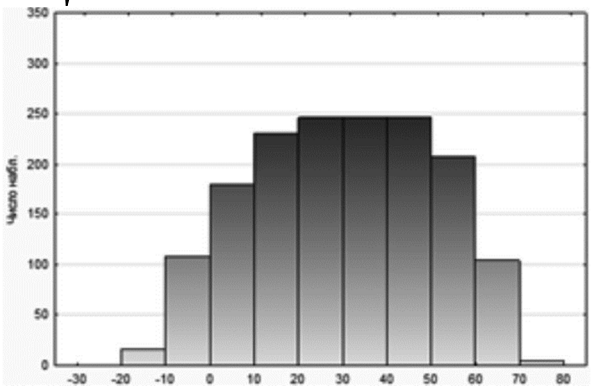

Рис. 3 а. Распределение проекции исходных данных на координатную ось «X»

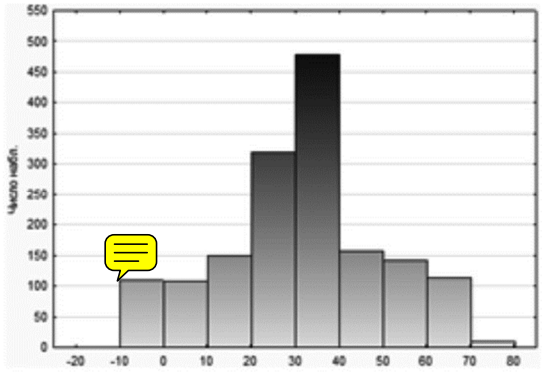

Рис. 3 б. Распределение проекции исходных данных на координатную ось «Y»

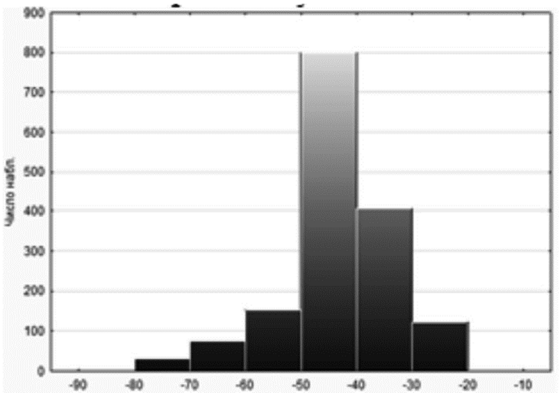

Рис. 4 а. Распределение проекции исходных данных на координатную ось «X′»

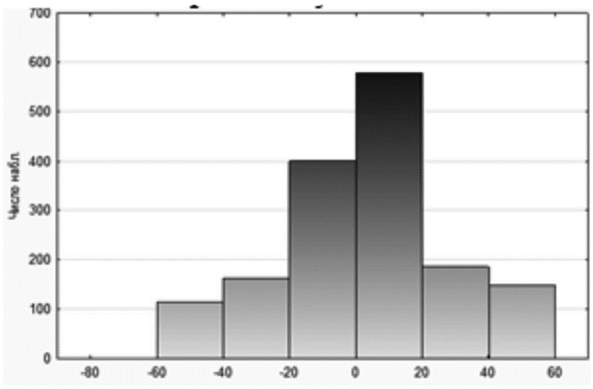

Рис. 4 б. Распределение проекции исходных данных на координатную ось «Y′»

Литература

1. A. Ortega P. Frossard J. Kovacevi ˇ c J. M. F. Moura and P. Vandergheynst. «Graph signal processing: Overview, challenges, and applications», Proc. IEEE, vol. 106, no. 5, pp. 808-828. May 2018.

2. Бериков В.Б. Выбор оптимальной сложности класса логических решающих функций в задачах анализа разнотипных данных: дис. д-р. тех. наук: 05.13.17. – Новосибирск, 2006. – 271 с.

3. Кручинин И.И. Моделирование и сравнительный анализ процессов распознавания и классификации многомерных объектов пересекающихся классов на основе представлений теории нечетких множеств и нейросетевых технологий: дис. канд. тех. наук: 05.13.17. – Калуга, 2003. – 209 с.

4. Sutha P., Jayanthi V. Fetal Electrocardiogram Extraction and Analysis Using Adaptive Noise Cancellation and Wavelet Transformation Techniques // J Med Syst. – 2018. № 21.

5. Белобродский В.А., Туровский Я.А., Вахтин А.А., Борзунов С.В., Кургалин С.Д. Обобщение метода цепочек локальных экстремумов для анализа сигналов различной природы // Цифровая обработка сигналов. – 2015. – № 1. – С. 35-38.

6. Пономарев А.В. Основы теории двумерной цифровой обработки сигналов в базисах Фурье с варьируемыми параметрами // Цифровая обработка сигналов. 2019. № 2. С. 12-20.

7. Утробин В.А. Методы обработки в условиях априорной неопределенности: дис. д-р. тех. наук: 05.13.17. – Нижний Новгород, 1997. – 410 с.

8. Коваленко А.П. Непараметрические методы анализа кластеров высокой плотности: дис. д-р. тех. наук: 05.13.17. – М., 1999. – 184 с.

9. Крестьянинова М.А. Применение методов контролируемой классификации для анализа биологических данных: дис. канд. физ.-мат. наук: 03.00.02: Москва, 2003. – 136 с.

10. M. Silveira, G. Figueiredo, R. Caputo New Time-Domain Approach for Digital Signal Processing: A Set of Experimental Measures for Systems with High Transmission Rates // Journal of Circuits, Systems and Computers. – 2019. Vol. 28 (05), P. 1950072.

11. Alpaydin E. Introduction to Machine Learning, MIT-Press, 2014. – 539 p.

12. Jolliffel. T. Principal Component Analysis, Series: Springer Series in Statistics, Springer, 2002. – 487 стр.

13. Le Roux B., Rouanet H. Geometric Data Analysis: From Correspondence Analysis to Structured Data, Springer Science & Business Media, 2014, pp. 297-332.

14. Лурье А. И. Аналитическая механика. – М.: Физматлит. – 1961. – 824 с.