КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 631−639

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

УДК: 004.02, 004.94

Метод представления дифракционных изображений XFEL для классификации, индексации и поиска

С. А. Бобков[1,a], А. Б. Теслюк[1,b], О. Ю. Горобцов[1,2], О. М. Ефанов[3], Р. П. Курта[2], В. А. Ильин[1,4], М. В. Голосова[1], И. А. Вартаньянц [2,5]

[1] Национальный исследовательский центр «Курчатовский институт»,
Россия, 123182, г. Москва, пл. Академика Курчатова, д. 1
[2]Немецкий электронный синхотрон ДЕЗИ,
Германия, D-22607, г. Гамбург, Ноткесштрассе, д. 85
[3]Научный центр лазеров на свободных электронах,
Германия, D-22607, г. Гамбург, Ноткесштрассе, д. 85
[4] Московский государственный университет им. М. В. Ломоносова,
Россия, 119991, г. Москва, ГСП-1, Ленинские горы, д. 1-52
[5]Национальный исследовательский ядерный университет «МИФИ»,
Россия, 115409, г. Москва, Каширское шоссе, д. 31

E-mail: [a]s.bobkov@grid.kiae.ru, [b]anthony.teslyuk@grid.kiae.ru

Получено 21 января 2015 г.

В работе представлены результаты применения алгоритмов машинного обучения: метода главных компонент и метода опорных векторов для классификации дифракционных изображений, полученных в экспериментах на лазерах на свободных электронах. Показана высокая эффективность применения такого подхода с использованием модельных данных дифракции лазерного пучка на капсиде аденовируса и вируса катаральной лихорадки, в которых учтены условия реального эксперимента на лазерах на свободных электронах, такие как шум и особенности используемых детекторов.

Ключевые слова: метод главных компонент, метод опорных векторов, когерентная визуализация

Вычисления выполнялись на компьютерных ресурсах ЦКП «Комплекс моделирования и обработки данных исследовательских установок мегакласса», поддерживаемого соглашением с Минобрнауки России о предоставлении субсидии № 14.621.21.0006.

КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 631−639

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

XFEL diffraction patterns representation method for classification, indexing and search

S. A. Bobkov[1], A. B. Teslyuk[1], O. Yu. Gorobtsov[1,2], O. M. Yefanov[3], R. P. Kurta[2], V. A. Ilyin[1,4], M. V. Golosova[1], I. A. Vartanyants [2,5]

[1] National Research Center "Kurchatov Institute", 1 Kurchatov Sq., Moscow 123182, Russia
[2]Deutsches Elektronen-Synchrotron DESY, 85 Notkestraße, D-22607 Hamburg, Germany
[3]Center for Free-Electron Laser Science, 85 Notkestraße, D-22607 Hamburg, Germany
[4] Lomonosov Moscow State University,GSP-1, 1-52 Leninskie Gory, Moscow, 119991, Russia
[5] National Research Nuclear University MEPhI, 31 Kashirskoe highway, 115409, Moscow, Russia

The paper presents the results of application of machine learning methods: principle component analysis and support vector machine for classification of diffraction images produced in experiments at free-electron lasers. High efficiency of this approach presented by application to simulated data of adenovirus capsid and bluetongue virus core. This dataset were simulated with taking into account the real conditions of the experiment on lasers free electrons such as noise and features of used detectors.

Keywords: principle component analysis, support vector machine, coherent diffraction imaging

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 3, pp. 631–639 (Russian).

I. Введение

При традиционных исследованиях малых биологических образцов с применением лазерного излучения возникают два ограничения. Во-первых, большинство белковых макромолекул не кристаллизуются, что препятствует использованию надежно отработанных методов рентгеновской кристаллографии. Во-вторых, для некристаллических образцов излучение вызывает повреждения, которые ограничивают возможное разрешение до нескольких десятков нанометров. Новый подход к визуализации с использованием одночастичной когерентной дифракции может позволить преодолеть второе ограничение и увеличить разрешение биологических объектов в субнанометровом диапазоне [Gaffney and Chapman, 2007; Seibert et al., 2011].

Идентичные образцы в случайной ориентации вводятся в луч лазера. Сами образцы разрушаются после выстрела за счет кулоновского взрыва [Neutze et al., 2000], но их дифракционные изображения регистрируются. Высокая мощность лазеров на свободных электронах (FELs) с фемтосекундными импульсами позволяет проводить эксперименты по определению структуры отдельных воспроизводимых частиц [Gaffney and Chapman, 2007; Mancuso, Yefanov, and Vartanyants, 2010]. Впоследствии по этим 2D-изображениям может быть восстановлена 3D-структура частицы. Данный метод уже был применен для исследования нанокристаллов [Chapman et al., 2011].

Тем не менее при таком подходе возникают дополнительные трудности. Одной из них является проблема ориентации частиц, т. к. положение, соответствующие каждому дифракционному изображению, изначально неизвестно. Правильные ориентации могут быть определены из множества измерений [Loh and Elser, 2009; Yefanov and Vartanyants, 2013; Fung et al., 2009] с учетом того факта, что все дифракционные изображения представляют собой сечения сферы Эвальда одинаковым трехмерным распределением интенсивности в обратном пространстве. Однако необходимо иметь достаточную выборку для того, чтобы определить ориентации изображений друг относительно друга, по крайней мере необходимо несколько сотен измерений. Это число может быть увеличено до нескольких тысяч, если учесть тот факт, что сигнал от одной макромолекулы является относительно слабым.

В процессе измерений возникает еще одна сложность. Не все получаемые изображения содержат дифракцию от образцов: большинство изображений пустые, а некоторые из них могут содержать, например, дифракцию от капель воды, несколько частиц некоторой примеси. Необходимо классифицировать изображения перед процедурой восстановления и использовать только те, которые соответствуют исследуемому образцу.

Можно выполнить классификацию вручную, но это занимает много времени. Недавно были предложены вычислительные методы для сортировки на основе метода главных компонент [Yoon et al., 2011].

В обработке изображений широко распространены различные алгоритмы определения характерных признаков. Они направлены на сокращение объемов данных, необходимых для точного описания исследуемых данных. Наиболее известной областью применения таких алгоритмов является машинное обучение: распознавание лиц [Yang et al., 2004], компьютерное зрение [Viola, Jones, 2001], лингвистика [Sebastiani, 2002] или интеллектуальный анализ данных [Berkhin, 2006].

Кроме того, существуют общие алгоритмы кластеризации, такие как метод главных компонент. Попытки применить эти алгоритмы непосредственно к дифракционным изображениям не принесли желаемых результатов. Для улучшения анализа дифракционных изображений требуется метод, который бы учитывал следующие составляющие дифракционной физики: процесс распространения лазерного импульса, характеристики луча, пространственные особенности молекул и т. д.

В этой статье представлен метод сортировки, основанный на сочетании метода главных компонент [Jolliffe, 2002] и корреляции угловой интенсивности дифракционных изображений [Altarelli, Kurta, and Vartanyants, 2010]. Для описания дифракционных изображений использовались характеристические векторы, которые вычисляются на основе угловой корреляции интенсивности. Связь между локальной структурой частиц и угловой корреляцией интенсивности была теоретически обоснована в недавней публикации М. Альтарелли и др. [Altarelli, Kurta and Vartanyants, 2010].

Наборы векторов для дифракционных изображений исследовались с помощью метода главных компонент. Сочетание двух подходов позволило выявить внутренние связи между дифракционными изображениями и добиться кластеризации данных в соответствии с исходными молекулами. Метод был разработан и проверен на основе данных моделирования, в которых учены условия реального эксперимента на лазерах на свободных электронах, такие как шум и особенности используемых детекторов.

II. Описание исходных данных

Для разработки метода был использован набор дифракционных изображений, полученный для молекул трех типов: капсид аденовируса [Zubieta, Blanchoin and Cusack, 2006], ядро вируса катаральной лихорадки (bluetongue virus core, 2BTV) [Grimes et al., 1998] и капля воды диаметром 10 нм.

Данные частицы имеют сравнимый размер, что усложняет классификацию. Капсид аденовируса обладает гексагональной симметрией, а 2BTV не обладает симметрией. Было сгенерировано по 1000 изображений для каждого типа частицы.

Весь набор модельных данных из 3000 изображений был случайно перемешан, затем из него был выделен обучающий набор. Контрольные данные о классификации полного набора изображений использовались только для сравнения итоговых результатов. Так как использовались данные моделирования, нельзя говорить о погрешности определения исходного типа молекул в контрольной классификации.

Размер обучающего набора должен удовлетворять двум условиям: содержать достаточную информацию для классификации полного набора и иметь как можно меньший размер, так как в экспериментальных условиях он составляется вручную. Варьируя размер обучающей выборки, мы установили, что для классификации полного набора данных из 3000 изображений с помощью метода опорных векторов достаточным является размер в 40 изображений. При меньшем размере появляются ошибки классификации, больший обучающий набор избыточен.

При моделировании использовались следующие параметры эксперимента: детектор установлен на расстоянии 100 мм от образца и имеет размеры 100×100 мм^2; разрешение 224×224 пикселя. Длина волны излучения равнялась 0.3 нм. Угловой размер спекла составлял 0.06 радиана.

Падающий луч имел гауссово распределение, ширина на полувысоте равнялась 150 нм, плотность потока — 107 фотонов/нм^2.

К изображениям был добавлен пуассоновский шум и beamstop с диаметром 10 пикселей. Моделирование проводилось в программе MOLTRANS, разработанной в DESY (Deutsches Elektronen-Synchrotron), которая учитывает особенности экспериментов по дифракции лазерного излучения на отдельных макромолекулярных объектах для лазеров на свободных электронах.

Дифракционные изображения, соответствующие разным образцам, показаны на рис. 1. Дифракционные изображения для капель воды заметно отличаются, тогда как изображения двух других типов сложно отделить друг от друга. Основная трудность — классифиировать BTV и аденовирус и показать, что различия в дифракционных изображениях могут быть использованы для успешной классификации.

III. Метод автоматической классификации дифракционных изображений

A. Характеристический вектор для дифракционных изображений

Для классификации нам необходимо извлечь из дифракционных изображений набор признаков, которые наилучшим образом связаны со структурой изучаемого объекта. Используя

факты о взаимосвязи спектра угловой кросскорреляционной функции интенсивности дифракционной картины и спектра электронной плотности, мы предлагаем в качестве характеристических векторов использовать спектр угловой автокорреляционной функции (1). В автокорреляционной функции содержится существенно меньшее количество информации, однако мы покажем, что этого достаточно для классификации изображений различных молекул.

$$C(q,\Delta)=\left\langle I(q,\phi)I(q,\phi+\Delta)\right\rangle_{\phi}. \tag{1}$$

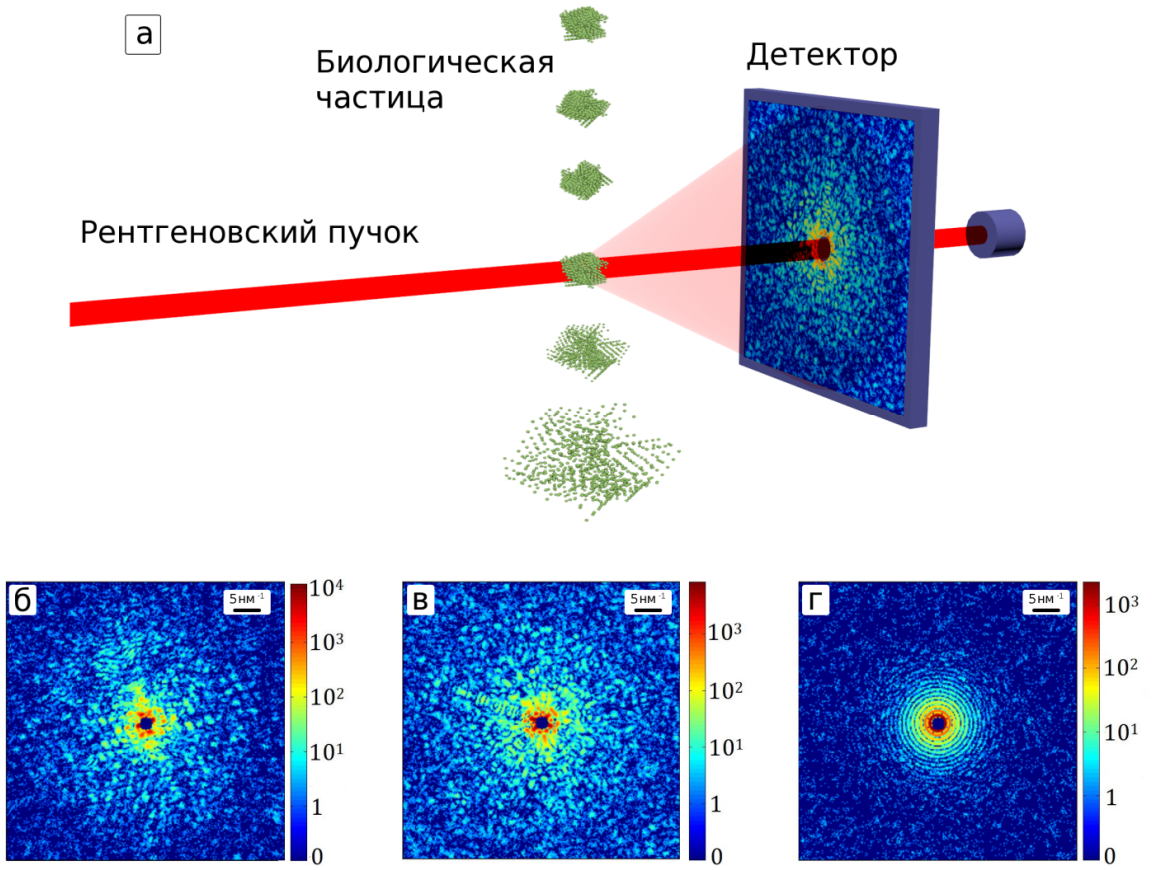

Рис. 1. (а) Схема эксперимента; (б), (в), (г) характерные дифракционные изображения различных молекул: (б) BTV; (в) аденовирус; (г) вода

Таким образом, каждому изображению будет соответствовать вектор:

$$F=\left\langle \bar{C}_q^1,\ldots,\bar{C}_q^n\right\rangle. \tag{2}$$

B. Кластеризация характеристических векторов с помощью метода главных компонент

Для классификации изображений мы уменьшаем размерность фазового пространства, определяемую размерностью характеристических векторов до плоскости, где точки, соответствующие отдельным дифракционным изображениям, могут быть визуально проанализированы и разделены, если возможно. Для этого мы применяем метод главных компонент (PCA).

PCA строит новый базис собственных векторов для матрицы ковариации данных, которые называются главными компонентами, и эти векторы упорядочены по убыванию соответствующих собственных значений. Затем ограниченное число первых главных компонент выбирается

как новое фазовое пространство. После этого анализируется проекция начального фазового пространства в новое пространство главных компонент.

Одна из особенностей PCA состоит в том, что пространство главных компонент, построенное с использованием первых n компонент, имеет максимальную дисперсию среди всех возможных ортонормальных базисов размерности n в фазовом пространстве. Таким образом, плоскость, построенная на основе первых двух главных компонент, будет содержать максимальную возможную дисперсию проекции данных.

Чтобы использовать метод главных компонент, мы строим матрицу: $A = \left\| a_{ij} \right\| = \left\| CCF_i^{(j)} \right\|$, где i соответствует изображению, а j — компоненте преобразования косинусов от угловой корреляции. Чтобы найти базис метода главных компонент, вычисляются собственные векторы матрицы ковариации данных. Самый быстрый способ сделать это — центрировать по столбцам матрицу A:

$$\overline{a}_{ij} = a_{ij} - \frac{1}{N}\sum_{k=0}^{N} a_{ki} \tag{3}$$

и затем, используя сингулярное разложение, записать $\left\| \overline{A} = \overline{a}_{ij} \right\|$ как

$$\overline{A} = U \Sigma V^T, \tag{4}$$

где U, V — унитарные матрицы, а Σ — диагональная. Легко показать, что столбцы матрицы V — это главные компоненты, которые мы ищем. Затем мы берем проекции для векторов, соответствующих каждом изображению и получаем координаты на плоскости (PC1, PC2).

Применяя метод главных компонент к набору характеристических векторов, мы получаем двумерную плоскость, где каждое дифракционное изображение описывается точкой. Если обработка изображений настроена правильно, разные типы изображений будут образовывать отдельные группы.

Если характеристические векторы, соответствующие разным группам, могут быть разделены в пространстве первых N главных компонент некоторым разделяющим правилом, то мы будем использовать это правило для классификации типов изображений.

Изображения, соответствующие каплям воды, группируются в одну точку, т. к. представляют собой концентрические окружности, и кросскорреляционная функция для таких изображений является константой, поэтому изображения для капель воды точно классифицируются по спектру кросскорреляционной функции. Результат применения метода главных компонент к набору модельных данных для аденовируса и 2BTV представлен на рис. 2. Видно, что дифракционные изображения образуют кластеры в многомерном пространстве. Данный результат подтверждает связь кросскорреляционных коэффициентов и пространственной структуры молекул, однако для классификации полного набора требуется более подходящий алгоритм.

C. Кластеризация характеристических векторов с помощью метода опорных векторов

Метод главных компонент (PCA) разделяет данные на основе среднего положения отдельных групп, а свойства данных требуют сфокусировать метод разделения на границе между группами. Этот подход был реализован с использованием линейного метода опорных векторов (Linear SVM).

SVM строит гиперплоскость в пространстве характеристических векторов высокой размерности, которая затем используется для разделения. Качество разделения достигается максимизацией расстояния от плоскости до ближайшей точки обучающего набора. Пусть есть обучающий набор (x_i, y_i), где x_i — это характеристический вектор и y_i — либо 1, либо –1 в зависимости от типа изображения, SVM находит решение проблемы оптимизации для всего обу-

чающего набора:

$$y_i\left(\mathbf{w}\cdot x_i-\mathbf{b}\right)\geq 1, \tag{5}$$

где ***w*** вектор нормали к гиперплоскости, а **b** определяет положение плоскости относительно начала координат вдоль вектора нормали **w**. Получившаяся гиперплоскость используется для классификации. Результат скалярного произведения характеристического вектора и **w** дает вероятность для частицы принадлежать определенному типу.

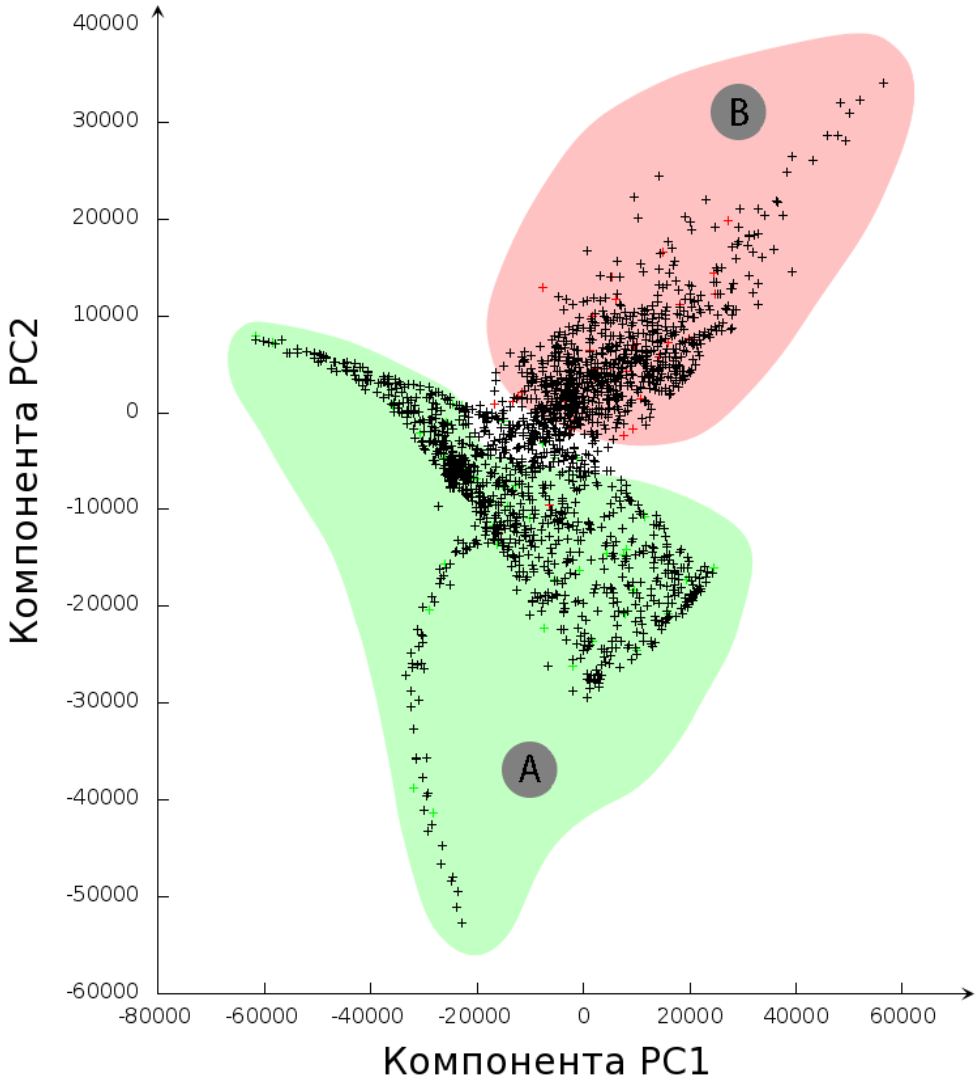

Рис. 2. Кластеризация модельных данных с помощью метода главных компонент. Каждому изображению соответсвует точка с координатами на плоскости первых двух главных компонент: PC1–PC2. Область A содежит изображения тренировочного набора, соответствующие только молекулам аденовируса, а область B — только молекулам 2BTV. Изображения между областями требуют более точной классификации

Существует набор различных расширений для метода опорных векторов, таких как мультиклассовый SVM или нелинейный SVM. Однако линейный SVM для двух классов был выбран как наиболее подходящий и имеющий наибольшее качество. Разделение для нескольких типов было реализовано последовательным применением метода опорных компонент к разделению определенного типа против всех остальных. Порядок выбора типа для разделения определяется сложностью отделения типа из априорных соображений.

Результат классификации набора модельных данных с помощью метода опортных векторов показан на рис. 3. При сравнении полученной классификации с точной классификаций набора модельных данных тип исходной молекулы был определен верно для 99 % дифракционных изображений.

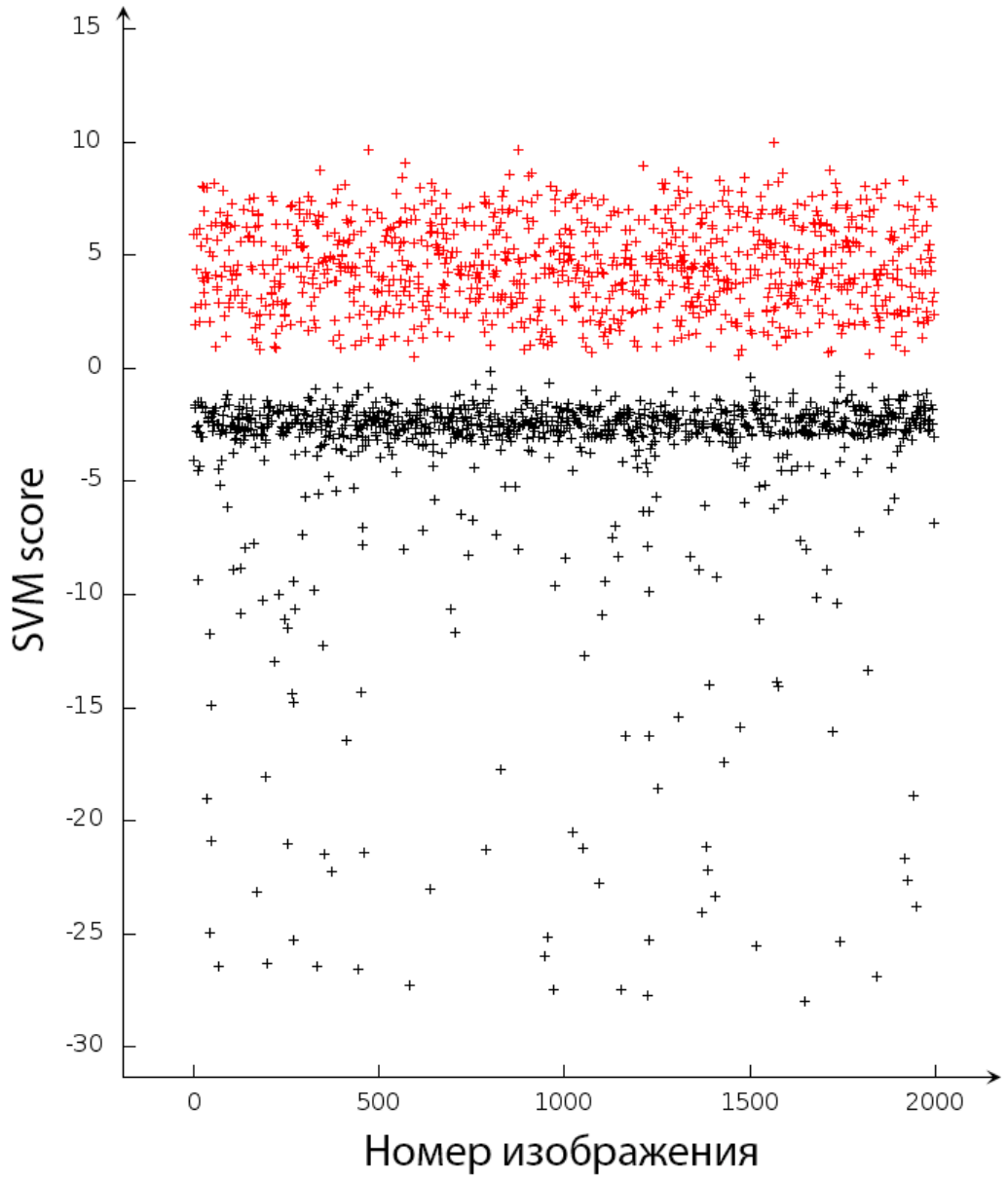

Рис. 3. Классификация на основе метода опорных векторов. Изображения аденовируса находятся в отрицательной части и имеют черный цвет, а изображения 2BTV — красный цвет. Параметр SVM score характеризует положение изображения относительно границы между двумя классами

Методы были реализованы с помощью языка Python, библиотеки матричных вычислений Numpy, математической библиотеки Intel Math Kernel Library, библиотеки Scikit-learn.

Высокая производительность вычислений достигалась с помощью применения технологии параллелизации вычислений OpenMP. Для вычислений использовался высокопроизводительный кластер центра коллективного пользования «Комплекс моделирования и обработки данных от исследовательских установок мегакласса» НИЦ «Курчатовский институт».

IV. Выводы

В работе мы представили два метода классификации дифракционных изображений на базе метода главных компонент (PCA) и на базе метода опорных векторов (BVM). Для модельных данных PCA позволяет показать кластеризуемость данных, однако не позволяет точно определить тип исходных молекул для всех изображений. Эффективность классификации SVM близка к 100 %. Превосходство SVM над PCA можно объяснить следующими соображениями: PCA ищет линейные комбинации свойств анализируемых объектов, которые наилучшим образом отличают все объекты друг от друга, в то время как SVM направлен на поиск оптимальной границы между двумя классами объектов, что больше соответствует задаче классификации классов изображений.

Наш метод классификации может быть применен в автоматизированной системе анализа данных, для поиска изображений, содержащих дифракционную картину исследуемых объектов, а также для индексирования наборов данных и поиска дифракционных изображений интересующих объектов. На его основе возможно построение самообучающихся алгоритмов классификации.

Список литературы

Altarelli M., Kurta R., and Vartanyants I. Physical Review B 82, 104207. — 2010.

Berkhin P. Grouping Multidimensional Data, Recent Advances in Clustering. — 2006. — P. 25–71.

Chapman H. N. et al. Femtosecond X-ray protein nanocrystallography // Nature. — 2011. — Vol. 470. — P. 73.

Fung R., Shneerson V., Saldin D. K., and Abbas O. // Nature Physics. — 2009. — Vol. 5. — P. 64.

Gaffney K. J. and Chapman H. N. Imaging atomic structure and dynamics with ultrafast X-ray scattering // Science. — 2007. — Vol. 316. — P. 1444.

Grimes J. M. et al. // Nature. — 1998. — Vol. 395. — P. 470.

Jolliffe I. T., ed. Principal Component Analysis. — Springer. — 2002.

Loh N.-T. D. and Elser V. // Phys. Rev. — 2009. — E 80, 026705.

Mancuso A. P., Yefanov O. M., and Vartanyants I. A. // J. Biotechnology. — 2010. — Vol. 149. — P. 229.

Neutze R., Wouts R., Van der Spoel D., Weckert E., and Hajdu J. Potential for biomolecular imaging with femtosecond X-ray pulses // Nature. — 2000. — Vol. 406. — P. 752.

Sebastiani F. ACM Computing Surveys (CSUR) Surveys 34, 1. — 2002.

Seibert M. M. et al. // Nature. — 2011. — Vol. 470. — P. 78.

Viola M., Jones P. Computer Vision and Pattern Recognition, CVPR. 1, I. — 2001.

Yang J., Zhang D., Frangi A., and Yang J.-Y. Pattern Analysis and Machine Intelligence. — 2004. — Vol. 26. — P. 131.

Yefanov O. M. and Vartanyants I. A. // J. Phys. B: At. Mol. Opt. Phys. — 2013. — 46, 164013.

Yoon C. H. et al. // Optics Express. — 2011. — Vol. 19. — 16542.

Zubieta C., Blanchoin L., and Cusack S. 273, 4336 — 2006.