P10-2012-123

Г. Е. Козлов *

# ИССЛЕДОВАНИЕ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ОТКЛИКОВ ДЕТЕКТОРОВ С ЯЧЕИСТОЙ СТРУКТУРОЙ

*E-mail: G.Kozlov@gsi.de

Козлов Г. Е. Р10-2012-123

Исследование алгоритмов кластеризации откликов детекторов с ячеистой структурой

Обсуждаются методы кластеризации и их применение для обработки откликов детекторов с ячеистой структурой, используемых в экспериментах физики высоких энергий. Описаны иерархические и итеративные методы. Рассматриваются преимущества, недостатки и применимость методов к конкретной задаче. Отобраны наиболее подходящие для решения данной задачи методы: Варда и одиночной связи. Выбор методов осуществлялся с учетом точности, эффективности и скорости работы каждого алгоритма. Разработан новый алгоритм кластеризации, учитывающий особенности задачи. Проведенные исследования алгоритмов на модельных данных показали, что разработанный алгоритм имеет лучшую производительность по сравнению со стандартными методами Варда и одиночной связи.

Работа выполнена в Лаборатории информационных технологий ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна, 2012

Kozlov G. E. P10-2012-123

Study of Clustering Algorithms for Detectors with the Pad Structure

We discuss clustering methods and their application to processing responses of the detectors with the pad structure that are used in high energy physics experiments. Hierarchical and iterative methods are described. We discuss their advantages, disadvantages and applicability to a particular task. Three parameters, accuracy, efficiency and speed, are used to characterize each method. The Ward and single linkage methods are currently considered to be the most appropriate choice for our problem. A novel clustering algorithm which takes into account the peculiarities of our task was developed. Simulation studies showed that the novel algorithm has the best performance as compared to the standard Ward and single linkage methods.

The investigation has been performed at the Laboratory of Information Technologies, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna, 2012

# ВВЕДЕНИЕ

Одним из важных проектов в физике высоких энергий (ФВЭ) на сегодняшний день является эксперимент CBM (Compressed Baryonic Matter) [1], который будет проводиться на строящемся в Дармштадте (Германия) ускорительном комплексе FAIR (Facility for Antiproton and Ion Research). В рамках данного эксперимента планируется изучение состояний ядерной материи, образующихся в ядро-ядерных соударениях при энергиях пучка 8–45 ГэВ/нуклон.

Получение данных осуществляется с помощью ряда детекторов, большинство из которых имеют ячеистую структуру, например: детектор переходного излучения TRD (Transition Radiation Detector), мюонный детектор MUCH (MUon CHamber). Частицы, проходя через координатные плоскости таких детекторов, в месте соприкосновения с детектирующей поверхностью вызывают срабатывание одной или нескольких ячеек. В таких условиях важным этапом обработки экспериментальных данных является максимально точное определение координат пролета каждой частицы. Для решения этой задачи требуется предварительная кластеризация данных с последующим определением точных координат центров найденных кластеров.

Частота поступления данных при проведении эксперимента CBM может достигать 10 МГц, при этом предполагается, что обработка экспериментальных данных должна вестись в темпе их поступления, т. е. в реальном времени. В таких условиях к алгоритмам кластеризации предъявляются повышенные требования по скорости работы. Кроме того, большое количество частиц (до нескольких тысяч), образующихся при каждом соударении, обуславливает высокую плотность и сложную топологию экспериментальных данных (сработавшие от близко летящих частиц ячейки могут соприкасаться), что серьезно осложняет их кластеризацию.

Решение возникающей задачи быстрой кластеризации данных является актуальной проблемой, имеющей существенное значение для повышения эффективности работы установок в экспериментах ФВЭ и требующей развития существующих и разработки новых алгоритмов, способных точно и эффективно проводить кластеризацию в столь сложных условиях.

## 1. ОСНОВНЫЕ НАПРАВЛЕНИЯ КЛАСТЕРИЗАЦИИ

В общем случае кластерный анализ — это задача разбиения заданной выборки объектов на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Кластер — группа элементов, характеризуемых общим свойством [2]. Перед началом кластеризации необходимо найти ту совокупность переменных, которая наилучшим образом отражает понятие сходства. В рамках данной задачи критериями для определения схожести являются координаты центров ячеек и амплитуды их зарядов.

Для проведения анализа данных используют разные меры сходства [3]. Выделяют четыре из них: коэффициент корреляции, меры расстояния, коэффициенты ассоциативности, вероятностные коэффициенты сходства [4]. Кластеры могут состоять из различных объектов, но все они обладают некоторыми свойствами, наиболее важными из которых являются плотность, дисперсия, размеры, форма и отделимость [2]. Кластеры могут быть перекрывающимися. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют спорными. Спорный объект — это объект, который по мере сходства может быть отнесен к нескольким кластерам. Неоднозначность данной задачи может быть устранена экспертом или аналитиком, но с учетом значимости скорости анализа данных при кластеризации откликов трековых детекторов неоднозначность должна быть разрешена непосредственно программой.

На сегодняшний день существует значительное количество методов кластеризации, различающихся подходом к обработке данных и областью применения. Рассмотрим подробнее семейства методов кластеризации, наиболее часто применяемые для обработки физических данных, а именно: иерархические агломеративные (объединяющие), дивизимные (разделяющие) и итеративные методы группировки. Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие. Итеративные методы основаны на перераспределении объектов среди заданного числа кластеров.

**Иерархические агломеративные методы (Agglomerative Nesting, AGNES)** Характеризуются последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров [5].

Работа таких методов связана с построением матрицы сходства размерностью $N \times N$ (где $N$ — число объектов), также называемой матрицей расстояний — квадратной матрицы типа «объект–объект» (порядка $N$), содержащей в качестве элементов расстояния между объектами в метрическом пространстве. В соответствии с этой матрицей на каждом шаге работы алгоритма наиболее схожие объекты объединяются, после чего меры сходства для образо-

вавшегося кластера пересчитываются. Всю последовательность объединений объектов можно представить в виде дендрограммы — древовидной диаграммы, каждая ветвь которой соответствует одному шагу работы алгоритма кластеризации. Это дерево изображает иерархическую организацию связей между несколькими элементами данных.

В общем виде алгоритм иерархического объединения объектов выглядит следующим образом:

1) рассчитывается матрица сходства между объектами;

2) определяются два наиболее близких кластера;

3) наиболее близкие кластеры объединяются;

4) алгоритм выполняется до тех пор, пока не будет остановлен или пока все объекты не будут объединены в один кластер.

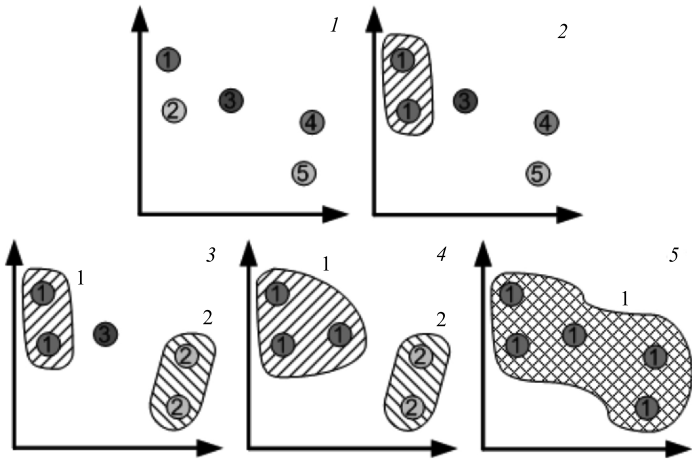


Рис. 1. Общая схема работы иерархических агломеративных методов

Для полной кластеризации такими методами на основе матрицы сходства размерностью $N \times N$ требуется $N - 1$ шагов. На первом шаге события (объекты) рассматриваются как самостоятельные кластеры. На последнем шаге все события объединяются в одну большую группу. Общая схема работы иерархических агломеративных методов приведена на рис. 1. Основные различия между иерархическими объединяющими методами связаны с выбором правил построения кластеров. Существует много различных правил группировки, каждое из которых порождает отдельный иерархический метод. Наиболее распространены четыре из них: одиночной связи, полной связи, средней связи и метод Варда.

Метод одиночной связи [2] осуществляет объединение по следующему правилу: объект будет присоединен к уже существующему кластеру, если

по крайней мере один из элементов кластера находится на том же уровне сходства, что и объект, претендующий на включение. Таким образом, присоединение определяется лишь наличием единственной связи между объектом и кластером. Результаты работы данного метода инвариантны к монотонным преобразованиям матрицы сходства, но как недостаток можно отметить склонность к образованию продолговатых кластеров-цепочек.

Метод полных связей [3] имеет более жесткие правила объединения, при которых для включения объекта в кластер требуется, чтобы сходство между ними было больше некоторого порогового уровня. Поэтому здесь имеется тенденция к обнаружению относительно компактных гиперсферических кластеров, образованных объектами с большим сходством.

Метод средней связи [2] проверяет среднее сходство рассматриваемого объекта со всеми объектами в уже существующем кластере, а затем, если найденное среднее значение сходства достигает или превосходит некоторый заданный пороговый уровень, объект присоединяется к этому кластеру.

Метод Варда [2] построен таким образом, чтобы оптимизировать минимальную дисперсию внутри кластеров. Эта целевая функция известна как внутригрупповая сумма квадратов или сумма квадратов отклонений (СКО):

$$\text{СКО} = x_j^2 - \frac{1}{n \cdot (-x_j)^2}, \tag{1}$$

где $x_j$ — значение признака $j$-го объекта. На первом шаге, когда каждый кластер состоит из одного объекта, СКО равна 0. По методу Варда объединяются те группы или объекты, для которых СКО получает минимальное приращение. Метод имеет тенденцию к нахождению или созданию кластеров приблизительно равных размеров и имеющих гиперсферическую форму.

**Иерархические дивизимные (делимые) методы (DIvisive ANAlysis, DIANA)** являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры [5]. Общая схема работы иерархических дивизимных методов приведена на рис. 2.

Существует два вида дивизимных методов: монотетический и политетический [6]. Монотетический кластер — это группа, все объекты которой имеют приблизительно одно и то же значение некоторого конкретного признака. Таким образом, монотетические кластеры определяются фиксированными признаками, определенные значения которых необходимы для принадлежности к кластерам. Политетические кластеры являются группами объектов, для принадлежности к которым требуется наличие определенных сочетаний из некоторого подмножества признаков.

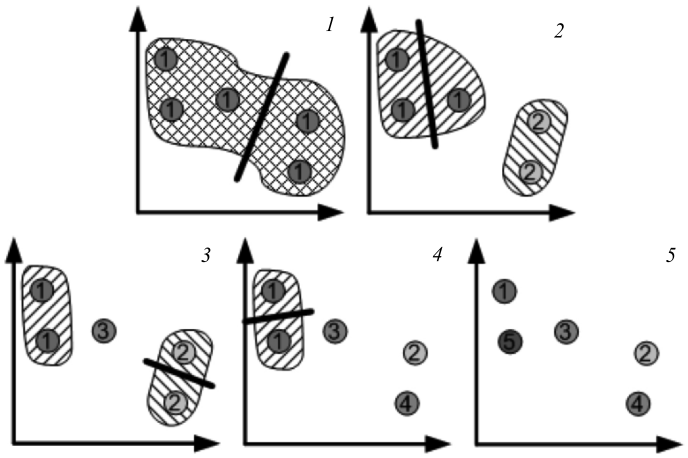


Рис. 2. Общая схема работы иерархических дивизимных методов

**Итеративные методы** группировки работают следующим образом [2]:

1) начать с исходного разбиения данных на некоторое заданное число кластеров; вычислить центры тяжести этих кластеров;

2) поместить каждую точку данных в кластер с ближайшим центром тяжести;

3) вычислить новые центры тяжести кластеров; кластеры не заменяются на новые до тех пор, пока не будут просмотрены полностью все данные;

4) шаги 2 и 3 повторяются до тех пор, пока не перестанут меняться кластеры.

Общая схема работы итеративных методов представлена на рис. 3. В отличие от иерархических агломеративных методов, которые требуют вычисления и хранения матрицы сходств между объектами размерностью $N \times N$, итеративные методы работают непосредственно с первичными данными. Поэтому с их помощью возможно обрабатывать довольно большие множества данных. Более того, итеративные методы делают несколько просмотров дан-

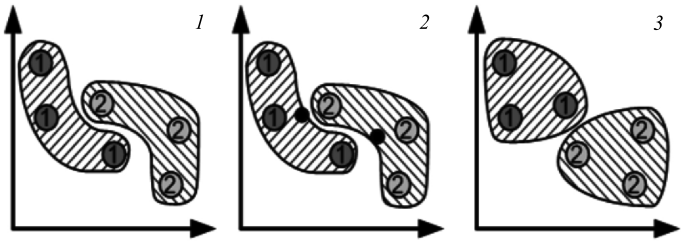


Рис. 3. Общая схема работы итеративных методов

ных и могут компенсировать последствия плохого исходного разбиения данных. Работа этих методов не может быть представлена в виде иерархического дерева, поскольку они порождают кластеры одного ранга. Большинство итеративных методов не допускает перекрытия кластеров.

Большинство свойств итеративных методов группировки могут быть описаны с помощью трех основных факторов: выбора исходного разбиения, типа итерации и статистического критерия. Различные комбинации этих факторов ведут к разработке методов, порождающих разные результаты при работе с одними и теми же данными.

Серьезным недостатком итеративных методов является проблема сходимости итераций не к оптимальному решению, достигаемому в минимуме используемого критерия, а к субоптимальному решению в некоем локальном минимуме. Это происходит, поскольку вычислительная реализация этих методов может выбрать лишь очень малую часть всех возможных разбиений, и возникает определенная вероятность, что будет выбрано субоптимальное разбиение. Поэтому такую проблему называют также проблемой локального оптимума. Еще одним недостатком итеративных методов является то, что число кластеров должно быть известно заранее, что не позволяет использовать их для кластеризации произвольных данных.

## 2. КЛАСТЕРИЗАЦИЯ ОТКЛИКОВ ТРЕКОВЫХ ДЕТЕКТОРОВ ЯЧЕИСТОЙ СТРУКТУРЫ

Задача кластеризации откликов трековых детекторов ячеистой структуры представляет собой разбиение данных, полученных в ходе эксперимента на группы — кластеры, и нахождение центров этих групп. Данные эксперимента представлены в виде набора ячеек на плоскости, обладающих собственными координатами, размерами и амплитудами полученной ими энергии. Объединение ячеек в группы осуществляется по следующим основным правилам:

1) В одну группу объединяется набор соседних ячеек. Объединение несвязанных ячеек недопустимо.

2) Каждая группа ячеек имеет свой глобальный максимум, наличие других локальных максимумов возможно только в случае пролета частицы под углом к детектирующей плоскости.

3) Форма образованных кластеров может быть различной: как круглой, так и вытянутой (если частица пролетает под углом к плоскости).

Количество частиц, прошедших через детектирующую плоскость, заранее неизвестно и должно быть определено в ходе работы алгоритма. Наиболее значимыми характеристиками алгоритмов кластеризации при исследовании результатов экспериментов в ФВЭ являются скорость, точность и эффективность. Следует также указать на зависимость этих характеристик

алгоритма от свойств самих кластеризуемых данных, таких как плотность потока частиц, степень зашумленности данных, диапазон разрядов, отводимых для регистрации амплитуд сигналов в ячейках, и др. Рассмотрим применимость описанных выше методов к задаче кластеризации откликов трековых детекторов ячеистой структуры.

Итеративные методы кластеризации требуют начального разбиения данных на определенное количество кластеров. Однако в рассматриваемой задаче количество частиц, прошедших через детектирующую плоскость, неизвестно заранее, поэтому для применения итеративных методов придется исследовать все возможные варианты разбиения данных. С учетом объема обрабатываемых данных (до нескольких сотен тысяч объектов) использование таких методов в данном случае нецелесообразно.

При использовании иерархических дивизимных методов большие кластеры делятся на меньшие. При таком подходе необходимо принимать решения о том, к какому из новых кластеров должна отойти та или иная ячейка. Это может привести к возникновению проблемы спорных объектов, то есть ячеек, которые могут быть присоединены к нескольким кластерам. Таким образом, применение иерархических дивизимных методов к решению рассматриваемой задачи возможно только при условии наличия быстрых алгоритмов обработки спорных объектов.

В иерархических агломеративных методах малые кластеры последовательно объединяются в более крупные и задавать заранее число кластеров не требуется. Для определения объединяемых кластеров используются меры расстояния. В этом случае образование спорных объектов маловероятно, поскольку требует совпадения минимальных мер расстояния для этих объектов. Таким образом, иерархические агломеративные методы больше прочих рассмотренных подходят для решения поставленной задачи. Из приведенных выше иерархических методов уместно будет рассмотреть методы Варда и одиночной связи. Обобщенный алгоритм работы этих методов приводился выше. Рассмотрим подробнее каждый из них и особенности их применения к нашей задаче.

Поскольку оба этих метода относятся к иерархическим объединяющим, в их основе лежит последовательное объединение объектов на основе матрицы сходства. Основные различия между методами Варда и одиночной связи связаны со способами вычисления расстояний между объектами и особенностями объединения объектов.

В методе Варда сходство между объектами (кластерами) определяется при помощи расстояния между их центрами масс. В рамках данной задачи расстояния для матрицы сходства вычисляются по следующей формуле:

$$d_{RS} = \frac{a_R(a_S)}{a_R + a_S} \cdot (\|\overline{X}_R - \overline{X}_S\|^2), \tag{2}$$

где $a_R$ и $a_S$ — амплитуды кластеров $R$ и $S$ соответственно; $\overline{X}_R$ и $\overline{X}_S$ — координаты центров масс кластеров; $\|\dots\|$ — евклидово расстояние. Для нашей задачи евклидово расстояние рассчитывается следующим образом:

$$E = \sqrt{(x_R - x_S)^2 + (y_R - y_S)^2}. \tag{3}$$

В результате формируется симметричная матрица. Затем определяются два кластера, расстояние $d_{RS}$ между которыми минимально. Эти кластеры объединяются, и вычисляется новый центр масс полученного кластера:

$$\overline{X}_R^{(\text{new})} = \frac{1}{a_R + a_S} \cdot (a_R \cdot \overline{X}_R + a_S \cdot \overline{X}_S). \tag{4}$$

После каждого объединения строка и столбец, соответствующие присоединенному кластеру, вычеркиваются из матрицы. Значения расстояний для образованного кластера пересчитываются по формуле (2) с учетом новых координат центра масс и амплитуды.

Метод одиночной связи предполагает, что для первичного формирования кластера производится объединение двух наиболее близких объектов. Далее к кластеру может быть присоединен тот объект, который имеет наименьшее расстояние хотя бы до одного из объектов, уже включенных в кластер. Таким образом, степень близости оценивается не между самими кластерами, а между принадлежащими им объектами. В рассматриваемой нами задаче степень близости между объектами можно определить по принципу расположения ячеек на детектирующей плоскости — ближайшими считаются соседние ячейки. В этом случае каждая обособленная группа ячеек с ненулевыми амплитудами образует отдельный кластер. Такой подход позволяет отказаться от вычисления матрицы сходства. Вместо этого мы использовали рекурсивную функцию для последовательного присоединения к ячейкам всех соседей с ненулевыми амплитудами. Данная функция имеет следующий алгоритм:

1) получить координаты ячейки $(X_1, Y_1)$ и/или номер ячейки $N_1$;

2) проверить список соседей ячейки $N_1$: если ячейка $N_n$ имеет амплитуду больше нуля, то запустить рекурсивную функцию, передав ей $(X_n, Y_n)$ и/или $N_n$;

3) завершить выполнение функции, подняться на более высокий уровень рекурсии.

Поскольку списки соседей строятся при инициализации данных о детекторе, это не замедляет работу алгоритма. Подобная реализация алгоритма быстра, но влечет за собой резкое падение эффективности при повышении количества близко расположенных кластеров.

Метод Варда способствует образованию кластеров круглой формы с минимальной дисперсией и способен показать достаточно высокую точность при обработке данных, но его работа связана с вычислением матрицы расстояний

размерностью $N \times N$, где $N$ — число объектов для анализа. Поскольку число таких объектов в экспериментальных данных может быть до нескольких сотен тысяч, обработка одного события может занимать значительное время (более секунды на событие), что при темпе поступления данных $10^7$ событий в секунду недопустимо.

Напротив, использование метода одиночной связи для решения данной задачи не связано с построением матрицы расстояний. Ближайшими соседями можно считать только те ячейки, которые непосредственно соприкасаются на плоскости детектора. Для объединения таких ячеек достаточно использовать последовательно объединяющую их рекурсивную функцию. Такой алгоритм достаточно быстр в исполнении, но влечет за собой резкое падение точности и эффективности при повышении количества близко расположенных кластеров. Таким образом, ни один из рассмотренных методов не решает поставленную задачу в полной мере. По этой причине остается потребность в разработке иных алгоритмов кластеризации.

Поскольку каждый из рассмотренных методов имеет ряд недостатков, нами был разработан специализированный алгоритм, учитывающий особенности поставленной задачи. В упрощенном виде этот алгоритм выглядит следующим образом:

1) сформировать первоначальный набор объектов: каждая ячейка с ненулевой амплитудой объявляется отдельным кластером;

2) рассмотреть все элементы первоначального набора объектов: если ячейка не является локальным максимумом, то присоединить ее к соседу с наибольшей амплитудой, иначе оставить без изменений.

Работа данного алгоритма на малом наборе данных (9 ячеек) проиллюстрирована на рис. 4. При инициализации данных для каждой ячейки с ненулевой амплитудой определяются ее характеристики: амплитуда и номер. Номер ячейки определяет, к какому кластеру она принадлежит. Если ячейка с номером A1 присоединяется к A2, то ее номер изменяется на A2. Таким образом, на первом шаге алгоритма каждая ячейка получает уникальный но-

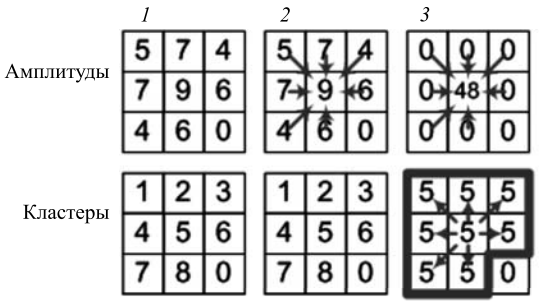


Рис. 4. Общая схема работы разработанного метода кластеризации

мер, а в конце работы все ячейки одного кластера имеют одинаковые номера. Изменение амплитуд ячеек в ходе кластеризации может вестись по различным правилам, в зависимости от реализуемого варианта алгоритма.

Работа такого алгоритма сходна с работой клеточного автомата [7]. Каждая ячейка представляет собой клетку автомата. «Живыми» считаются те ячейки, амплитуда которых больше нуля. Исходные данные, представляющие собой набор «живых» клеток, запускаются на эволюцию. В конце эволюции все ячейки каждого кластера имеют одинаковые номера, характерные только для данного кластера, т. е. процесс кластеризации считается завершенным. В зависимости от подхода к объединению ячеек эволюция может завершаться за одну или несколько эпох.

Такой алгоритм можно отнести как к иерархическим агломеративным, так и к итеративным методам кластеризации. Сходство с агломеративными методами состоит в том, что в начале кластеризации каждый объект (ячейка) считается отдельным кластером, а в ходе эволюции они объединяются в более крупные. Но сам процесс объединения идет по правилам, схожим с правилами итеративных методов.

Представленную схему работы алгоритма можно рассматривать в двух вариантах в зависимости от числа процессоров, используемых для обработки данных: однопоточном и многопоточном. При работе в однопоточном режиме целесообразно перестраивать кластеры (т. е. изменять номера ячеек присоединяемого кластера) после каждого присоединения. В таком случае, каждый последующий шаг алгоритма может зависеть от результатов предыдущих. Процесс кластеризации полностью завершится после однократной обработки каждого из первоначальных объектов. Использование для перестройки кластеров рекурсивной функции, объединяющей соседние ячейки с заданным номером кластера, позволяет производить эту операцию значительно быстрее полного перебора ячеек.

В многопоточном режиме работы объекты или их группы должны обрабатываться одновременно, а не последовательно, как в однопоточном. В таком случае при анализе объектов формируется только список связей объединяемых кластеров, а их перестройка производится после обработки всех ячеек. В настоящей работе рассматривается однопоточный вариант исполнения разработанного алгоритма.

## 3. ИССЛЕДОВАНИЕ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ

Поскольку нами был разработан собственный алгоритм кластеризации откликов трековых детекторов ячеистой структуры, его характеристики (точность, скорость, эффективность) должны быть определены на некоторых входных данных. Эти характеристики следует сравнить с аналогичными ве-

личинами существующих методов (метод Варда, метод одиночной связи). По результатам сравнения можно судить об успешности разработки метода кластеризации и целесообразности его использования в ФВЭ.

Для исследования свойств алгоритма необходимо провести ряд численных экспериментов. Так как для реальных данных точное расположение кластеров и координаты их центров неизвестны, необходимо подготовить достаточно большую выборку модельных данных с заранее известными параметрами. После проведения кластеризации эти параметры оцениваются статистическим путем для определения точности, эффективности и скорости работы каждого из сравниваемых методов кластеризации. Для обеспечения статистической достоверности и приемлемого расчетного времени мы использовали следующие параметры моделирования. Размер поля $200 \times 200$ точек. Размер ячейки $8 \times 8$ точек. Масштаб: 1 точка $= 1$ мм. В ходе эксперимента проводится генерация данных и их обработка каждым из рассматриваемых алгоритмов. Цикл генерации и обработки данных в дальнейшем обозначается как событие. Эксперимент состоит из 100 событий. В ходе каждого события проводятся следующие действия:

1) на заданном поле в случайном порядке расставляются несколько кластеров;

2) заполненное поле анализируется при помощи исследуемых алгоритмов;

3) полученные результаты сохраняются для обработки.

Каждый кластер состоит из некоторого числа граничащих ячеек с ненулевыми значениями амплитуд. Для моделирования кластера генерируется 1000 точек, отклонение от центра для каждой из которых распределено по нормальному закону со средним ноль и среднеквадратичным отклонением 10. Затем определяется, в какие ячейки попали эти точки, так что значение амплитуды ячейки равно числу попавших в нее точек. Попадание точки в некоторую ячейку определяется в соответствии с ее координатами. При таких условиях образуются достаточно крупные кластеры (40–50 ячеек) круглой формы. Это позволяет более качественно, чем на малых кластерах, оценить возможности алгоритмов по разделению близко расположенных объектов. На рис. 5 представлен пример сгенерированного события.

Главной задачей исследования является нахождение кластеров как связанных групп ячеек и определение их центров, указывающих на место прохождения частицы через детектирующую плоскость. Однако в процессе обнаружения кластеров могут возникнуть ошибки, связанные с потерей некоторых «слабых» или слившихся с другими кластеров, а также с нахождением фиктивных кластеров, образованных случайными скоплениями точек. Фиктивные кластеры образуются в том случае, если возникают ложные локальные максимумы амплитуд при слишком малом размере ячейки относительно размера кластера. Эти ошибки процедуры поиска кластеров в свою очередь вызывают искажения в координатах центров найденных кластеров. Таким

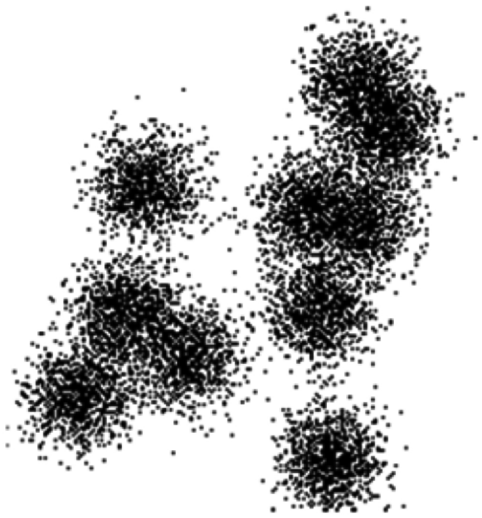

Рис. 5. Пример смоделированного события

образом, мы можем определить следующие критерии качества алгоритма поиска кластеров:

1) точность определения координат $X$ и $Y$ центра кластера, которая получается статистическим путем по большому числу событий как среднеквадратичное значение отклонения координат каждого из найденных кластеров от его значения, заложенного в модель;

2) эффективность алгоритма, определяемая как отношение общего числа найденных к числу смоделированных центров кластеров по всем событиям;

3) скорость работы алгоритма;

4) количество фиктивных (ошибочно найденных) кластеров.

Предлагаемый численный эксперимент позволяет определить эти характеристики качества алгоритма в зависимости от количества кластеров, а также сравнить соответствующие характеристики разработанного нами алгоритма, методов Варда и одиночной связи.

В соответствии с разработанной моделью эксперимента нами проведено исследование свойств рассматриваемых алгоритмов. Результаты данного исследования представлены в таблице. В графе «Время работы» указано суммарное время обработки алгоритмом всех 100 событий в эксперименте. Тестирование проводилось на компьютере с процессором Intel Xeon x3210, 2 Гб оперативной памяти. Для наглядности сравнения методов кластеризации полученные данные представлены в виде графиков на рис. 6–8.

**Результаты исследования свойств алгоритмов**

| Число кластеров | Ошибка по $X$, мм | Ошибка по $Y$, мм | Эффективность, % | Фиктивные кластеры, шт. | Время работы, с |
|---|---|---|---|---|---|
| Разработанный алгоритм | | | | | |
| 1 | 1,000 | 1,190 | 100 | 0 | 0,039 |
| 3 | 1,370 | 0,740 | 100 | 0 | 0,045 |
| 5 | 1,254 | 1,190 | 100 | 0 | 0,079 |
| 7 | 1,459 | 1,391 | 95 | 0 | 0,123 |
| 10 | 1,850 | 2,055 | 94 | 5 | 0,193 |
| 15 | 2,437 | 2,533 | 85 | 9 | 0,373 |
| 20 | 2,647 | 2,565 | 80 | 5 | 0,578 |
| Метод Варда | | | | | |
| 1 | 1,120 | 1,380 | 100 | 0 | 0,128 |
| 3 | 1,180 | 1,090 | 100 | 0 | 1,175 |
| 5 | 1,320 | 1,600 | 100 | 0 | 4,707 |
| 7 | 2,011 | 1,911 | 97 | 0 | 7,919 |
| 10 | 2,303 | 2,232 | 97 | 11 | 21,775 |
| 15 | 2,995 | 2,855 | 91 | 15 | 45,138 |
| 20 | 3,200 | 3,000 | 85 | 8 | 85,750 |
| Метод одиночной связи | | | | | |
| 1 | 1,000 | 1,190 | 100 | 0 | 0,021 |
| 3 | 1,714 | 1,995 | 93 | 0 | 0,025 |
| 5 | 3,743 | 4,567 | 73 | 0 | 0,047 |
| 7 | 4,308 | 4,786 | 53 | 0 | 0,098 |
| 10 | 4,683 | 4,846 | 45 | 0 | 0,153 |
| 15 | 5,086 | 5,500 | 38 | 0 | 0,268 |
| 20 | 6,335 | 6,145 | 17 | 0 | 0,325 |

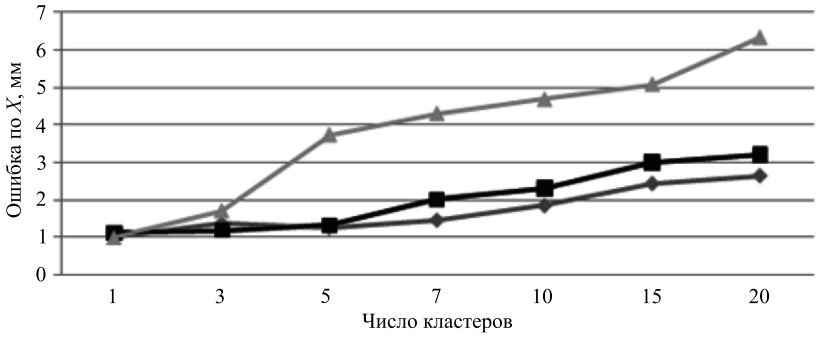


Рис. 6. График зависимости точности от числа кластеров: ◆ — разработанный алгоритм; ■ — метод Варда; ▲ — метод одиночной связи

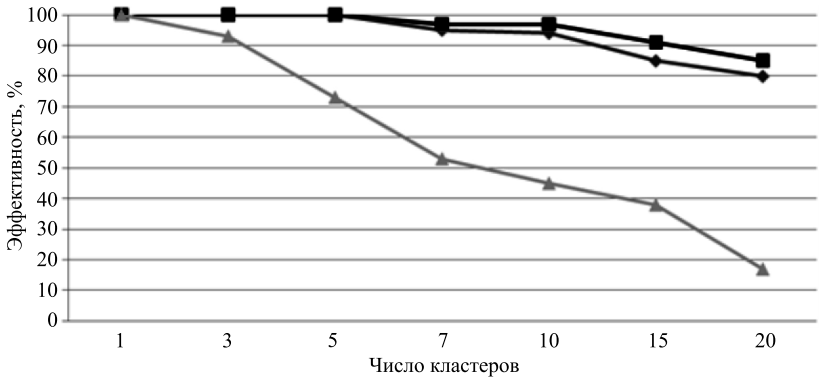


Рис. 7. График зависимости эффективности от числа кластеров; обозначения, как на рис. 6

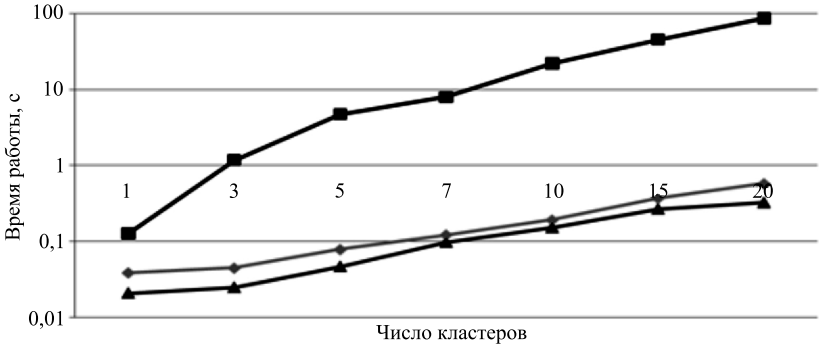


Рис. 8. График зависимости скорости работы от числа кластеров; обозначения, как на рис. 6

Как видно из таблицы и графиков, при работе с модельными данными лучшие результаты показал разработанный нами алгоритм. Худшие результаты показал метод одиночной связи, поскольку он стремится к образованию цепочек связанных кластеров и не имеет механизмов их разделения. Но в силу своей простоты этот метод работает быстрее остальных. Для метода одиночной связи характерны хорошая точность (ошибка до 2 мм) и эффективность (более 95 %) при малом количестве кластеров, но с возрастанием этой величины точность и эффективность стремительно снижаются. При числе кластеров более 5 для данного эксперимента использование метода одиночной связи невозможно, поскольку его эффективность падает ниже 70 %.

Метод Варда показал неплохие результаты (ошибка — до 4 мм, эффективность — от 85 %) даже при максимальном количестве кластеров. Основным недостатком данного метода является слишком долгая работа (до 85 с на 100 событий). Это связано с необходимостью построения матрицы расстояний и ее частичного пересчета в процессе работы. Такая скорость работы является недопустимой для нашей задачи.

Разработанный нами алгоритм показал наилучшие результаты среди рассматриваемых. Он обладает наилучшей точностью (ошибка до 2,6 мм) вне зависимости от числа кластеров. Отставание от метода Варда по эффективности можно считать в пределах методической погрешности. Разработанный алгоритм серьезно превосходит метод Варда по скорости работы (более 100 раз при большом числе кластеров), но отстает от метода одиночной связи. Таким образом, тестирование разработанного алгоритма кластеризации можно считать успешным.

Приведенная модель эксперимента позволяет провести исследование характеристик алгоритмов для общего случая кластеризации. При этом все кластеры имеют сходную суммарную амплитуду и форму. В отличие от таких модельных данных данные эксперимента CBM могут содержать кластеры различных размеров и форм, в зависимости от особенностей пролета частицы через плоскость детектора. По этой причине необходимо рассмотреть особенности работы алгоритмов кластеризации с данными, максимально приближенными к реальным. Для этого нами использованы данные, смоделированные в программной оболочке эксперимента CBM — CBMROOT [8] с учетом реальных условий эксперимента CBM для первого слоя первой станции детектора MuCH. В качестве входных данных используется часть детектирующей плоскости размером $37 \times 37$ ячеек. Ячейки имеют квадратную форму, сторона — 4 мм. Рассматриваемая область находится близко к центру плоскости, что обуславливает достаточно сложную топологию расположения кластеров. На рис. 9 приведены результы кластеризации таких данных каждым из рассматриваемых алгоритмов.

По результатам кластеризации разработанным методом обнаружено 90 кластеров, методом Варда — 102 кластера, методом одиночной связи — 42 кластера. Методом одиночной связи верно определены только отдельные, не связанные друг с другом кластеры. Таким образом, можно сказать, что данный метод не справился с обработкой данных в полной мере. Разработанный алгоритм и метод Варда показали близкие результаты, различающиеся лишь в некоторых аспектах. Метод Варда, использующий при объединении матрицу расстояний, распознал наибольшее количество кластеров. При этом он показал себя лучше в тех случаях, когда более слабый кластер соприкасается со значительно более сильным (имеются в виду амплитуды соприкасающихся ячеек). Однако в случаях продолговатых кластеров метод Варда тяготеет к их разбиению на множество малых, по 1–3 ячейки в каждом.

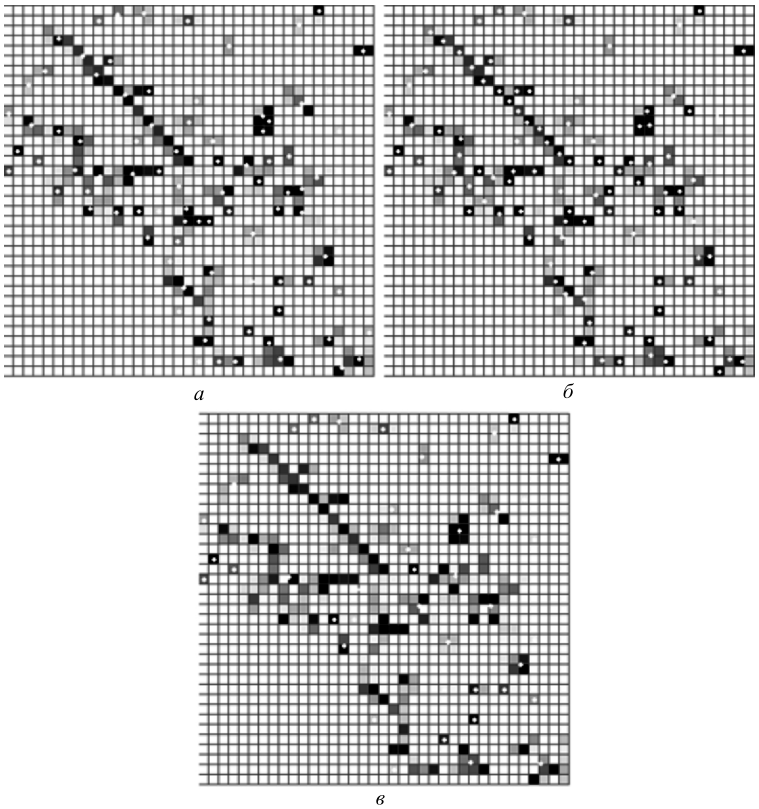


Рис. 9. Тестирование алгоритмов на данных, смоделированных в программной оболочке эксперимента СВМ: *а*) разработанным алгоритмом; *б*) методом Варда; *в*) методом одиночной связи

Разработанный алгоритм при кластеризации опирается на следующее правило: каждой ячейке с локальным максимумом амплитуды соответствует один кластер. При таком подходе лучше определяются продолговатые кластеры, имеющие один локальный максимум, но существует вероятность ошибочного присоединения слабого кластера к более сильному. Фактически, такой кластер имеет один локальный максимум, но представляет собой две зрительно отделимые группы ячеек. Вследствие этого возможно небольшое падение эффективности в сравнении с методом Варда. На решении данной проблемы следует сосредоточить внимание при работе над алгоритмом в будущем.

## ЗАКЛЮЧЕНИЕ

Настоящая работа посвящена вопросу выбора наилучшего алгоритма для решения задачи кластеризации откликов трековых детекторов ячеистой структуры в экспериментах ФВЭ. Рассмотрев наиболее популярные методы кластеризации, мы выбрали среди них наиболее пригодные для решения поставленной задачи. Основным критерием при выборе явились возможности методов для работы с данными определенного формата. Результаты эксперимента представляются в виде набора данных, содержащих координаты ячеек на детектирующей плоскости и их амплитуды. Задача кластеризации сводится к разделению ячеек с ненулевыми амплитудами на заранее не известное число кластеров.

Рассмотрев основные преимущества и недостатки, мы отобрали наиболее подходящие, но обладающие противоположными характеристиками методы:

1) метод Варда, имеющий высокую точность и эффективность, но обладающий низкой скоростью за счет необходимости построения матрицы расстояний;

2) метод одиночной связи, имеющий высокую скорость за счет простоты алгоритма, но с низкими точностью и эффективностью.

Поскольку стандартные методы кластеризации имеют определенные недостатки, которые не позволяют использовать их в полной мере для кластеризации откликов трековых детекторов ячеистой структуры, нами был разработан специализированный алгоритм кластеризации для решения рассматриваемой задачи.

Для определения характеристик нового алгоритма и сравнения его с существующими мы разработали модель и провели ряд численных экспериментов. В данном исследовании алгоритмы, реализованные на языке программирования C++, запускались для обработки одинаковых входных данных, после чего сравнивались их точность, эффективность и скорость работы.

По результатам исследования разработку нового алгоритма можно признать успешной, поскольку он не уступает методам Варда и одиночной связи по основным характеристикам и не имеет явных недостатков, помимо возможности ошибочного объединения соседних кластеров при определенных условиях. Дальнейшую работу над этим алгоритмом следует сосредоточить на решении указанной проблемы, а также реализации параллельного варианта работы в режиме многопоточности. Это позволит серьезно уменьшить время и повысить точность и эффективность анализа данных.

Хочу поблагодарить А. А. Лебедева за полезные дискуссии при подготовке данной работы. Благодарю проф. Г. А. Ососкова за постановку задачи кластеризации данных и плодотворные обсуждения во время моей работы. Выражаю благодарность Е. Л. Крышеню за помощь в предоставлении данных для тестирования алгоритмов.

## ЛИТЕРАТУРА


1. Compressed Baryonic Matter Experiment — Technical Status Report. Collaboration CBM, January 15, 2005.
2. *Мандель И. Д.* Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.
3. *Дюран Б., Оделл П.* Кластерный анализ: Пер. с англ. Е. З. Демиденко / Под ред. А. Я. Боярского. М.: Статистика, 1977. 128 с.
4. *Гмурман В. Е.* Теория вероятностей и математическая статистика: Учебное пособие для вузов. М.: Высшая школа, 2004. 479 с.
5. *Чубукова И. А.* Интеллектуальный анализ данных (Data Mining). Лекция 13. 2006. URL: http://www.intuit.ru/department/database/datamining/13/2.html.
6. *Anderberg M. R.* Cluster Analysis for Applications. N. Y.: Academic Press, 1973. 359 p.
7. *Ососков Г. А., Тихоненко Е. А.* Новый генератор случайных чисел на базе двумерного клеточного автомата // Математическое моделирование. 1996. Т. 8, вып. 12. С. 77–84.
8. *Kryshen E.* Simulation of MUCH Detector and Readout Issues // 17th CBM Collaboration Meeting. Dresden, April 4–8, 2011.


Получено 23 ноября 2012 г.

Редактор *М. И. Зарубина*

Подписано в печать 16.01.2013.
Формат 60 × 90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 1,31. Уч.-изд. л. 1,6. Тираж 225 экз. Заказ № 57887.

Издательский отдел Объединенного института ядерных исследований
141980, г. Дубна, Московская обл., ул. Жолио-Кюри, 6.
E-mail: publish@jinr.ru
www.jinr.ru/publish/