Д.Е. МОЗОХИН,
В.А. КАЛЯГИН

Сравнительный анализ алгоритмов кластеризации в сетях фондовых рынков

УДК 519.24

ФГАОУ ВПО
«Национальный
исследовательский
университет
«Высшая
школа экономики»,
г. Нижний Новгород

В работе рассматриваются различные алгоритмы кластеризации в сетях фондовых рынков. Основное внимание уделяется вопросу качества кластеризации и интерпретации полученных результатов. Показано, что модифицированный алгоритм минимального остовного дерева является наиболее адекватным алгоритмом кластеризации активов фондового рынка.

Работа выполнена при поддержке гранта РНФ 14-41-00039.

Целью работы является исследование алгоритмов кластеризации в применении к анализу сетей фондовых рынков. В качестве входных данных используются доходности активов фондового рынка. Задача кластеризации заключается в нахождении разбиения множества всех активов так, чтобы каждый кластер состоит из объектов, близких по некоторой метрике, а объекты из других кластеров существенно отличались. Используемые в работе методы исследования основаны на анализе, сравнении и обобщении теоретических подходов, сборе информации с финансовых рынков, а также их обработке. Работа является одной из немногих попыток изучения рынков с помощью кластерного анализа.

Меры близости и расстояния

При решении задачи кластеризации необходимо зафиксировать метрику ρ, с помощью которой будет производиться расчет

расстояния между объектами. В силу специфики поставленной задачи, особый интерес представляют меры, ориентированные на их применение к временным рядам.

1. Евклидово расстояние – классическая метрика, представляющая собой геометрическое расстояние между объектами в многомерном пространстве. Предположим, что имеется два ряда длины n: $Q = q_1, q_2, \dots, q_n$ и $C = c_1, c_2, \dots, c_n$, тогда данная мера вычисляется по формуле:

$$ED(Q, C) = \sqrt{\sum_{i=1}^{n} (q_i - c_i)^2}$$

К преимуществам данной меры можно отнести низкую сложность вычисления (O(n)) и простоту интерпретации.

2. Линейная корреляция Пирсона – характеристика существования линейной зависимости между двумя величинами. Для определенных выше Q и C коэффициент корреляции вычисляется по формуле:

$$\rho(Q, C) = \frac{\sum_{i=1}^{n} (q_i - \bar{q})(c_i - \bar{c})}{\sqrt{\sum_{i=1}^{n} (q_i - \bar{q})^2 \sum_{i=1}^{n} (c_i - \bar{c})^2}},$$

где $\bar{q} = \frac{1}{n} \sum_{i=1}^{n} q_i$, $\bar{c} = \frac{1}{n} \sum_{i=1}^{n} c_i$ – выборочные средние, $\rho(Q, C) \in [-1, 1]$.

Отметим, что коэффициент корреляции определяет расстояние:

$$D(Q, C) = \sqrt{2(1 - \rho(Q, C))},$$

3. Динамическая трансформация временной шкалы (Dynamic time warping, DTW) – мера для вычисления расстояния между временными рядами. Эта метрика предоставляет возможность нелинейного сопоставления двух временных рядов за счет минимизации расстояния между ними. Существенным отличием является возможность ее вычисления для временных рядов разной длины, т.е. для $Q = q_1, q_2, \dots, q_n$ и $C = c_1, c_2, \dots, c_m$, где в общем случае $n \neq m$.

На первом этапе вычисления строится матрица расстояний d, размера $n \times m$, которая содержит расстояния (чаще всего

используется Евклидова метрика) между двумя точками q_i и c_j, где $i = \overline{1,n}$; $j = \overline{1,m}$.

На втором этапе вычисляют матрицу трансформация D, элементы которой находятся из следующего соотношения:

$D_{ij} = d_{ij} + \min(D_{i-1\,j}, D_{i-1\,j-1}, D_{i\,j-1})$.

На третьем шаге вводят понятие пути трансформации, обозначаемого W. W- набор соседних элементов матрицы D, представляющий собой путь, минимизирующий расстояние между временными рядами Q и C. $W = w_1, w_2, \ldots, w_K$, где $\max(n,m) \leq K \leq m+n-1$. На путь трансформации накладывается ряд условий:

- Граничные условия: $w_{11} = d_{11}$ и $w_{nm} = d_{nm}$. Это условие нужно для того, чтобы путь покрыл все точки временных рядов.

- Условие монотонности: любые два соседних элемента пути W, $w_k = d_{ij}$ и $w_{k-1} = d_{i-1\,j-1}$, удовлетворяют неравенствам: $w_i - w_{i-1} \geq 0$ и $w_j - w_{j-1} \geq 0$. Данное ограничение способствует тому, что путь продвигается вперед или остается на месте.

- Условие непрерывности: любые два соседних элемента пути W, $w_k = d_{ij}$ и $w_{k-1} = d_{i-1\,j-1}$, удовлетворяют неравенствам: $w_i - w_{i-1} \leq 1$ и $w_j - w_{j-1} \leq 1$. Данное ограничение гарантирует, что каждый индекс пути в матрице d увеличится не больше чем на единицу.

Финальным этапом является вычисление значения метрики:

$$dtw(Q,C) = \frac{\min(\sum_{i=1}^{K} d(w_i))}{K}$$

Поскольку алгоритм перебирает все клетки матрицы трансформации, то сложность алгоритма составляет $O(nm)$ (см.[9]).

Алгоритмы кластеризации

Для проведения наиболее полного анализа были рассмотрены алгоритмы, принципы работы которых существенно отличаются.

1. Метод к-средних (k-means) – алгоритм разделительной кластеризации, принцип которого основан на итеративном

разбиении входных данных на предопределенное количество кластеров *k*. Этот алгоритм является одним из самых популярных методов кластеризации, поскольку он отличается своей простотой и низкой вычислительной сложностью.

В процессе работы алгоритма минимизируется суммарное отклонение элементов кластеров от их центров:

$$V = \sum_{i=1}^{k} \sum_{c_j \in C_i} d(c_j, C_i^0),$$

где $C = C_1, C_1, \ldots, C_k$ – разбиение на *k* кластеров, C_i^0- центр *i-го* кластера, *d* – используемая метрика.

Алгоритм включает в себя следующую последовательность шагов:

- Определение количества кластеров k;

- Инициализируются k начальных центров путем произвольного выбора элементов из исходной выборки;

- Для каждого элемента из множества входных данных выбирается ближайший центр кластера, таким образом, происходит процесс формирования начальных кластеров;

- Пересчитываются центры тяжести кластеров, в которые и происходит смещение имеющихся центров. Новый центр представляет собой вектор, элементы которого являются средними значениями признаков, вычисленных по всем объектам соответствующего кластера.

Последние два шага повторяются итеративно, изменяя при этом границы и центры кластеров. Условием остановки алгоритма является устойчивость разбиения. Алгоритм обладает быстрой сходимостью. Дэвидом и Васильвицким [1] было показано, что сложность алгоритма равна $2^{\Omega(\sqrt{n})}$, где *n* - мощность входных данных.

2. Алгоритм кластеризации на основе минимального остовного дерева дерева (Minimum spanning tree, MST) – один из методов кластеризации на графах, вершинам которых соответствуют элементы выборки, а ребрам – расстояния между парами объектов. К преимуществам таких подходов относят наглядность, простоту и возможность внесения улучшений на основе геометрических соображений.

Минимальное остовное дерево – остовное дерево графа минимального веса. Для вычисления минимального остовного дерева используется алгоритм Краскала:

• Упорядочиваем по не убыванию веса всех ребер;

Пока множество ребер непустое:

• Последовательно соединяем вершины, которые инцидентны ребрам с минимальным весом и при этом не образуют цикл;

Для разбиения на *k* кластеров по алгоритму MST удаляем (*k-1)* ребро с максимальными весами из остовного дерева. Полученный лес, состоящий из *k* меньших деревьев, образует искомые кластеры.

Алгоритмам К-средних и MST необходимо явно задавать число кластеров. Методы, которые будут рассмотрены далее лишены данного недостатка и способны автоматически определять мощность оптимального разбиения.

3. Кластеризации планарного максимально отфильтрованного графа (Planar maximally filtered graph, PMFG) .

В работе [13] рассмотрено обобщение минимального остовного дерева и предложена новая структура, названная *PMFG.* Максимально отфильтрованный граф – взвешенный максимальный планарный граф, добавление любого ребра к которому, нарушает свойство планарности. Для построения таких графов используют модифицированный алгоритм Краскала:

• Упорядочивание по не убыванию весов всех ребер;

Пока множество ребер непустое:

• Фиксируем ребро;

• Если добавление ребра не нарушает планарности графа, то соединением вершины, которые инцидентны фиксированному ребру.

К преимуществам *PMFG* относят способность сохранить больше информации сети, чем минимальное остовное дерево, поскольку они содержат $3(N-2)$ *и* $(N-1)$ ребер соответственно, где N – число вершин. Также, одно из важных свойств *PMFG* – он содержит *MTS* в качестве подграфа.

Алгоритмом, позволяющим выполнить кластеризацию планарного максимально отфильтрованного графа является Directed Bubble Hierarchical Tree (DBHT), описанный в [12]. Шаги алгоритма состоят в следующем:

Шаг 1: Вводим понятие разделяющей клики размера 3. Это такие клики, удаление из графа вершин которых приводит к образованию двух несвязанных частей, которые могут быть соединены только удаленной кликой. Объединение этих двух частей и клики снова дают максимально отфильтрованный планарный граф. Вложенность клик образует иерархию. Процесс разбиения производится до тех пор, пока не переберутся все клики. Результатом является набор планарных графов, называемых “bubbles”, которые связаны друг с другом через разделяющие клики. Направленность дерева ассоциируется с ребрами $b_i b_j$ путем сравнения сумм весов ребер в этих компонентах. Направление указывает на вершину с большим суммарным весом. Листья этого дерева определяются как ключевые структуры, обладающие сильной связью, и их рассматривают в роли центров кластеров. Вершина b_i, соединенная направленным путем в дереве с листом b_α – относится к кластеру α.

Шаг 2: После шага 1 возможна ситуация, когда вершина b_i принадлежит нескольким кластерам. Для того чтобы получить дискретную кластеризацию, необходимо однозначно приписать каждую вершину исходного графа к кластерам. Это достигается после выполнения двух шагов.

1. Расчет силы притяжения вершины v исходного графа к листьям “bubble” дерева.

$$X(v, b_\alpha) = \frac{\sum_{u \in V(b_\alpha)} w_{vu}}{3(|V(b_\alpha)-2|)},$$

где w_{vu} – вес ребра (v,u). Вершина v приписывается к той вершине “bubble” дерева, сила притяжения с которой больше. После этого, каждая вершина b_α имеет набор уникальных вершин, образующих множество $V^0(\alpha)$.

2. Перераспределение оставшихся вершин (v_5, v_7, v_9) с помощью вычисления среднего кротчайшего пути:

$$\bar{L}(v, \alpha) = mean\{l(v,u) | u \in V^0(\alpha) \cap v \in V(h_\alpha)\},$$

где $l(v,u)$ - длина кротчайшего расстояния между вершинами v и u, h_α - поддерево “bubble” дерева, относящееся к кластеру α. Вершина v приписывается к кластеру, с которым имеем наименьшее среднее кратчайшее расстояние. В результате образуется дискретное разбиение вершин $V(G)$ на подмножества $V(\alpha)$, $V(\beta)$, …

Вычислительная сложность *DBHT* алгоритма - $O(|V^3|)$, где $|V|$ – число вершин графа (количество переменных в исходной выборке).

4. Алгоритмы, основанные на поиске оптимального значения функции модулярности. Модулярность(Modularity) – мера качества кластеризации. Методы кластеризации, основанные на её максимизации - являются одними из самых популярных среди алгоритмов, позволяющих автоматически определять количество кластеров [2], [3], [8]. Задача таких методов – поиск оптимального разбиения, максимизирующего значение функции модулярности. В целом, задачи такого рода являются NP-трудными, поэтому было разработано множество эвристических подходов. В рамках данной работы будет использован жадный агломерационный алгоритм, предложенный одним из авторов функции модулярности [10]. Рассматриваемый метод имеет следующий принцип:

1. Каждая вершина помещается в отдельный кластер;

2. Итеративно объединяются сообщества и отбираются на каждом шаге только те объединения, которые дают наибольший вклад в модулярность. После слияния *i-го* и *j-го* кластеров, значение модулярности изменится на $\Delta Q = 2(e_{ij} - a_i a_j)$, которое вычисляется за константное время;

Описанный выше алгоритм выполняется за время $O((m+n)n)$ и выполняется для различного числа разбиений.

Индексы оценки качества кластеризации

В результате применения алгоритмов кластеризации необходимо оценить качество полученных разбиений. Для этого используют индексы оценки качества. Для анализа были выбраны три индекса качества, имеющие различную природу.

1. Индекс Дэвиса-Болдина (Davies-Bouldin index, DB) – внутренняя схема оценки качества разбиения на основе количественных характеристик данных [4].

Формально DB имеет следующий вид:

$$DB = \frac{1}{m}\sum_{i=1}^{m} \max_{i \neq j}\left\{\frac{d(C_i)+ d(C_j)}{d(C_i, C_j)}\right\},$$

Где $d(C_i) = \frac{1}{m_i}\sum_{j=1}^{m_i} d(x, C_i^0)$ – внутрикластерное расстояние; m_i – количество элементов в кластере i; C_i^0 – центр кластера. d – заданная мера, $d(C_i, C_j)$ – межкластерное расстояние. Оно может вычисляться как расстояние между самыми близкими или наиболее удаленными точками кластеров *i* и *j*. Также, широкое распространение получило представление межкластерного расстояния с помощью вычисления удаленности центров кластеров.

Заметим, что DB является безразмерным и принимает неотрицательные значения. В отличие DI, наименьшее значение индекса DB соответствует лучшему разбиению данных на кластеры.

2. Кластерно-векторный баланс (Clustered-vector balance, CVB) – внутренний индекс оценки, основанный на кластерном балансе. Главное отличие от *CB* состоит в определении межкластерного расстояния, которое в *CVB* не учитывает расстояния до глобального центра [11].

Пусть $\Lambda = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{m_i} d(c_j^i,\ C_i^0)$ – внутрикластерное расстояние, которое соответствует среднему расстоянию между элементами кластеров и их центрами. $\Gamma = \frac{1}{k(k-1)}\sum_{i=1}^{m}\frac{n_i}{n}\sum_{j\neq i}^{m} d(C_i^0, C_j^0)$ – межкластерное расстояние, выражающее среднее расстояние между центрами кластеров. Очевидно, что лучшее решение задачи кластеризации должно максимизировать следующую функцию:

$$CVB = \Gamma - \Lambda,$$

поскольку оно описывает ситуацию, когда компактные кластеры являются хорошо разделимыми.

3. Модулярность(Modularity) - мера качества кластеризации, на основе строится широкий класс алгоритмов, оптимизирующих ее значение.

Впервые этот индекс был предложен в [10], где авторы описали новый класс алгоритмов кластеризации на графах. Общая идея заключалась в итеративном вычислении меры, основанной на количестве путей, проходящих через каждое ребро графа и удаление ребер с максимальным значением этой меры. Для оценки качества разбиений было введено понятие модулярности.

Пусть $\exists k$ кластеров и матрица E размером $k \times k$, в которой e_{ij}- доля ребер между *i-ой* и *j-ой* компонентами. Тогда след матрицы E определяет суммарную долю ребер внутри кластеров.

$$Trace\ E = Tr\ E = \sum_{i=1}^{k} e_{ii}$$

Очевидно, что чем больше след, тем лучше качество разбиения, поскольку большая часть ребер попадет в кластеры, которые будут слабо связаны между собой. Однако полностью нельзя опираться на этот показатель, так как при размещении всех ребер в одном кластере $Tr\ E = 1$, но это не дает никакой информации о структуре данных.

Обозначим за a_i долю ребер между *i-м* кластером и остальными. Тогда эта величина имеет вид: $a_i = \sum_{j=1}^{k} e_{ij}$. Если d_i- суммарная мощность всех вершин кластера i, а L- общее число ребер в графе, то $a_i = \frac{d_i}{L}$. На основе сделанных обозначений, функция модулярности имеет следующий вид:

$$Q = \sum_{i=1}^{k} (e_{ii} - a_i^2)$$

Интуитивно понятно, что если значение e_{ii} значительно больше a_i^2, то это означает, что существует много ребер в *i-ой* компоненте, чем бы мы ожидали в произвольном графе и это действительно кластер. Заметим, что количество вершин сети и внутри кластеров не влияют на значения Q. Данная мера качества позволяет оценить долю ребер, попавших в кластер с учетом вычета ожидаемого значения такой же величины, но в случае произвольного графа с таким же числом компонент. Нетрудно заметить, что наибольшее значение функции модулярности соответствует лучшему

разбиению. Что касается области принимаемых значений, они изменяются на отрезке от 0 до 1. По словам авторов, на практике значения $Q \in [0.3;0.7]$ свидетельствуют о наличии хорошей кластерной структуры.

Модулярность является достаточно новой характеристикой, которую сейчас активно изучают. Существует большое количество работ, посвященных этой теме. Например, в [2] обсуждаются свойства модулярности и приводится ряд свойств алгоритмов кластеризации, основанных на максимизации модулярности.

Модулярность имеет и недостатки. Главным является вычислительная сложность алгоритмов, основанных на оптимизации функции модулярности. Кроме того, модулярность имеет некоторое ограничение (resolution limit), которое влияет на обнаружение кластеров. В статье [6] исследуется наличие у модулярности внутреннего масштаба (scale), кластеры, меньшие которого, могут быть не обнаружены. Стоит понимать, что максимизация значений Q может привести к тому, что результирующее разбиение может являться комбинацией меньших кластеров оптимальной структуры.

Сравнение алгоритмов кластерного анализа в сетях фондовых рынков

Для проведения кластерного анализа использованы данные с финансовых рынков стран БРИКС: Бразилия, Россия, Индия и Китай. Некоторые результаты по сетевому анализу рынков других стран представлены в [5]. Главным критерием отбора финансовых активов являлся уровень их капитализации. Было отобрано следующее число активов 166 (Россия), 108 (Китай), 127 (Индия), 66 (Бразилия). Период наблюдений – 1 год, с 01.01.2013 до 31.12.2013.

Для сравнения алгоритмов необходимо было проанализировать зависимость качества кластеризации от выбранной метрики и метода. Другим важным аспектом качества является возможность содержательной интерпретации кластерных структур. Можно предположить, что полученные в ходе кластерного анализа разбиения будут ассоциированы с различным индустриальным секторам. Данная гипотеза может быть обусловлена тем, что на активы из одного сектора действуют одинаковые внешние факторы,

такие как экономическая обстановка, информационные потоки, политика и другие.

Анализ данных с финансовых рынков начнем с разбиений на два кластера. Результаты представлены в Таблице 1.

В таблице содержатся значения индексов оценки качества разбиений, полученных применением различных методов и мер близости. Ячейка (i, j) выделена красным цветом в том случае, если оптимальное значение индекса i соответствует методу j. Зеленым цветом указаны размеры образованных кластеров. Пустые значения в таблицах для методов *Modularity* и *DBHT* означают, что количество кластеров, полученных после применения соответствующего алгоритма не совпадает с предустановленным. Заметим, что при разбиении на 2 кластера, метод, основанный на MST, является лучшим с точки зрения валидационных индексов Дэвиса-Болдина и кластерно-векторного баланса. Важно отметить, что при использовании этого метода, функция модулярности, выступающая в роли индекса – голосует за метод К-средних, что выглядит разумнее. Особенно этот факт заметен для метрики *DTW.*

При разбиении данных на большее количество кластеров замечания, сделанные выше, сохраняют свою актуальность. Также, было замечено, что при увеличении количества кластеров кластерно-векторный баланс имеет оптимальное значение для разбиений, полученных методом максимизации функции модулярности. Например, этот факт отражен в случае использования корреляции Пирсона на китайском рынке. Однако, в этих же условиях, индекс Дэвиса-Болдина голосует за кластеризацию, имеющую несколько кластеров единичного размера, полученную на основе *MST*. О её не оптимальности свидетельствует эвристической метод максимизации модулярности, для которого *DB* индекс отличается на 0.04, но найденные кластеры заметно больше. При рассмотрении результатов для различных выборок мы можем сделать вывод о том, что предпочтительнее использовать метод кластеризации на основе функции модулярности, если главным критерием оценки найденных кластерных структур являются значения валидационных индексов.

Таблица 1

Значения индексов валидации при разбиении данных на два кластера.

Data-set: Russia (166 x 261)				
Metric: Euclidean	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	19174,6904746542[34;122]	0,124476015042624[165;1]	[166]	[70;85;42;3;2]
Clustering vector balance	-0,590537516	1,710181154		
Modularity	0,00360967	0,000165679		
Metric: Pearson	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	2,05435125526302[126;40]	1,63958429512401[165;1]	[75;64;27]	[28;1;22;1;1;1;1;1;1;1;1;1;75;1;1;12;2;25]
Clustering vector balance	-0,710905918	-0,545499837		
Modularity	0,242151666	0,003594945		
Metric: DTW	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	160471,167363508[53;113]	0,110434902098569[165;1]	[166]	[166]
Clustering vector balance	-4,737688833	16,02332015		
Modularity	0,434678473	0,011975613		
Data-set: China (108 x 261)				
Metric: Euclidean	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	2,27031411192833[68;40]	0,203970559672708[107;1]	[21;23;7;8;20;29]	[37;38;1;3;1;1;1;1;1;24]
Clustering vector balance	-0,593079023	1,783055816		
Modularity	0,140954236	0,001723023		
Metric: Pearson	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	3,86197850017772[54;54]	1,4417038382386[107;1]	[20;14;14;31;29]	[22;51;1;28;6]
Clustering vector balance	-0,834020358	-0,334627326		
Modularity	0,151411247	0,001103308		
Metric: DTW	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	42830,8466316125[49;59]	0,258602296869018[107;1]	2,97857481055348[76;32]	[108]
Clustering vector balance	-8,621769459	7,92498067	-5,672566434	
Modularity	0,495713306	0,018347051	0,417009602	
Data-set: India (127 x 261)				
Metric: Euclidean	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	4758,10522702995[51;76]	0,149298108190858[126;1]	2,60946181009824[107;20]	3,76067038654226[67;60]
Clustering vector balance	-0,376586851	0,848898002	-0,230817482	-0,266019606
Modularity	0,011026847	0,000207973	0,005908521	0,011446043
Metric: Pearson	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	2,10782038150425[118;9]	1,2148447650297[125;2]	3,44540351114975[91;36]	[26;5;26;21;20;16;2;2;3;4;2]
Clustering vector balance	-0,711737586	-0,345558525	-0,750427249	
Modularity	0,017714808	0,003064443	0,062849177	
Metric: DTW	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	47552,9175717751[29;98]	0,472508664144095[126;1]	[127]	[127]
Clustering vector balance	-2,974697619	0,195150843		
Modularity	0,352408705	0,015624031		
Data-set: Brazil (66 x 260)				
Metric: Euclidean	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	705,824855827008[51;15]	0,20500219595621[65;1]	1,78516076416649[54;12]	[14;28;14;4;6]
Clustering vector balance	-0,496202123	0,681969816	-0,176704743	
Modularity	0,020790085	0,000926633	0,018789097	
Metric: Pearson	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	2,15880854829802[29;37]	1,50407837221783[65;1]	[26;6;19;16]	[1;22;12;18;10;3]
Clustering vector balance	-0,689917183	-0,38609932		
Modularity	0,330040625	0,005431347		
Metric: DTW	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	21023,633199247[34;32]	0,320999161394727[65;1]	2,12711117294194[57;9]	[66]
Clustering vector balance	-3,60030951	2,003070426	-1,534286718	
Modularity	0,499540863	0,029843893	0,23553719	

Таблица 2

Значение индексов оценки качества для разбиений активов китайского финансового рынка на 5 кластеров

Data-set: China (108 x 261)				
Metric: Pearson	K-means	MST	DBHT (PMFG)	Modularity
Davies-Bouldin	3,05935495925807[9;40;19;11;29]	1,44535992810461[104;1;1;1;1]	1,84387030388024[20;14;14;31;29]	1,48210594791544[22;51;1;28;6]
Clustering vector balance	-1,167218024	-0,745208332	-0,549397703	-0,523378685
Modularity	0,128022866	0,004409575	0,198552177	0,227554267

Анализируя зависимость качества кластеризации на основе различных метрик важно понимать, что сравнение можно производить только для индексов оценки качества, которые является безразмерными величинами. В нашем случае - это функция модулярности, значения которой максимизируются в результате применения *DTW* метрики. Также - индекс Дэвиса-Болдина, выделяющий Евклидово расстояние и *DTW* метрику. Аналогичный вывод был получен и при кластеризации шаблонов потребления электроэнергии [7], где авторы показали, что расстояние Евклида, обладающее низкой вычислительной сложностью и простой интерпретацией, способно предоставить хорошее общее решение. Несмотря на сложность метрики DTW, ее применение также оправдывает себя, поскольку способно улучшить качество решения задачи кластеризации.

Изучение феномена MST

Проведя большое количество экспериментов на различных рынках мы установили, что разбиение на основе минимального остовного дерева имеет одну большую компоненту и несколько единичных кластеров. Для того, чтобы получить более содержательную информацию предлагается выбрать некоторое пороговое значение и удалить все ребра, имеющие больший вес. После чего, находим разбиение, игнорируя все кластеры, которые меньше порогового значения. Например, уровнем отсечения может быть средний вес ребер, а размер наименьшего кластера ограничен тремя вершинами.

Согласно предложенной стратегии, для данных с бразильского рынка при использовании корреляции Пирсона, мы получаем три кластера изображенных на Рис.6.

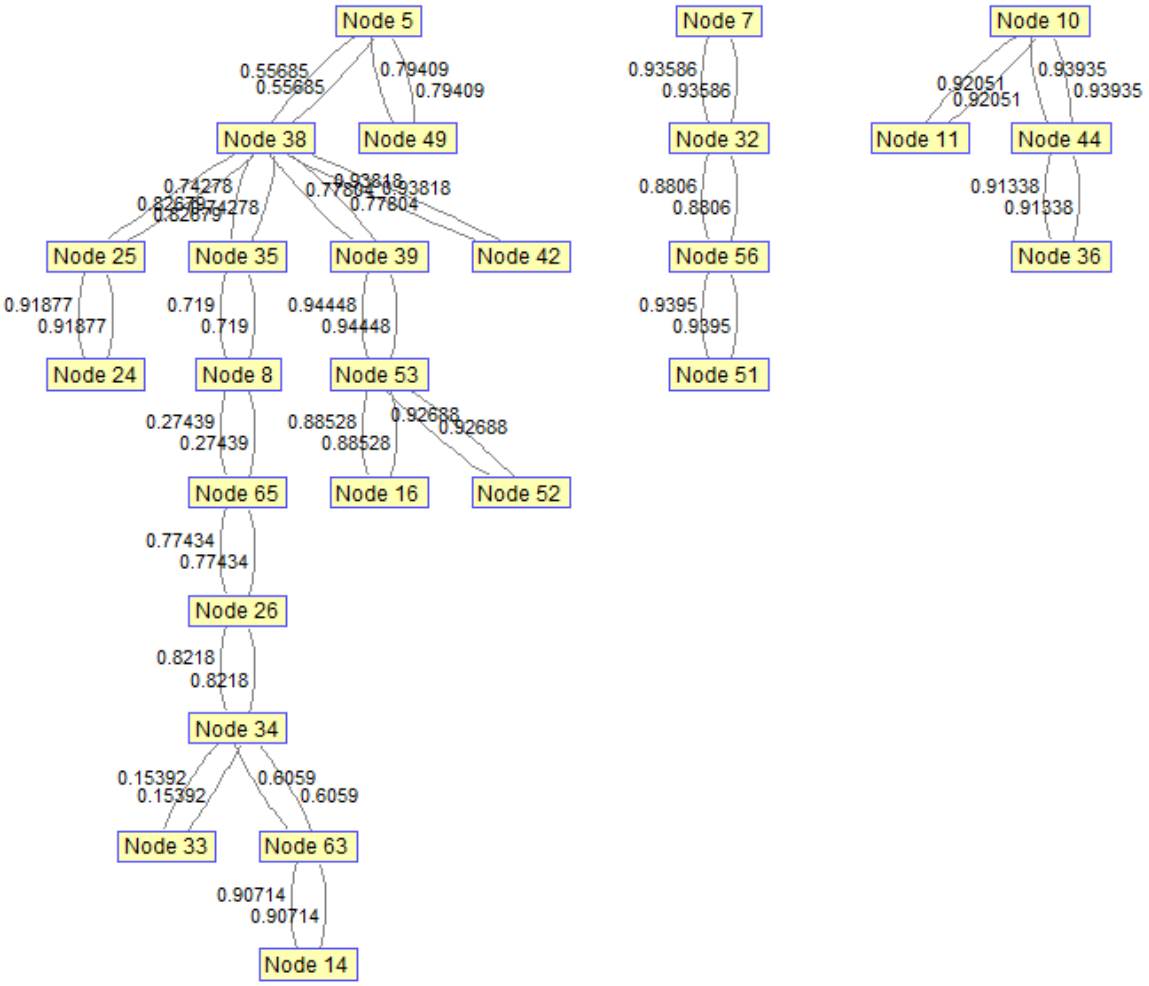

Рис. 1. Кластеризация данных с бразильского рынка при помощи минимального остовного дерева.

В результате образовались действительно содержательные кластеры с точки зрения интерпретации. Для более подробной интерпретации результатов кластерного анализа рассмотрим рынок Бразилии. С помощью индустриального классификатора (Industry Classification Benchmark, ICB) определим, к какому из десяти секторов экономки принадлежит каждый актив. Применение модифицированного метода MST дает разбиения активов в кластерах по секторам, представленные на рисунке 2.

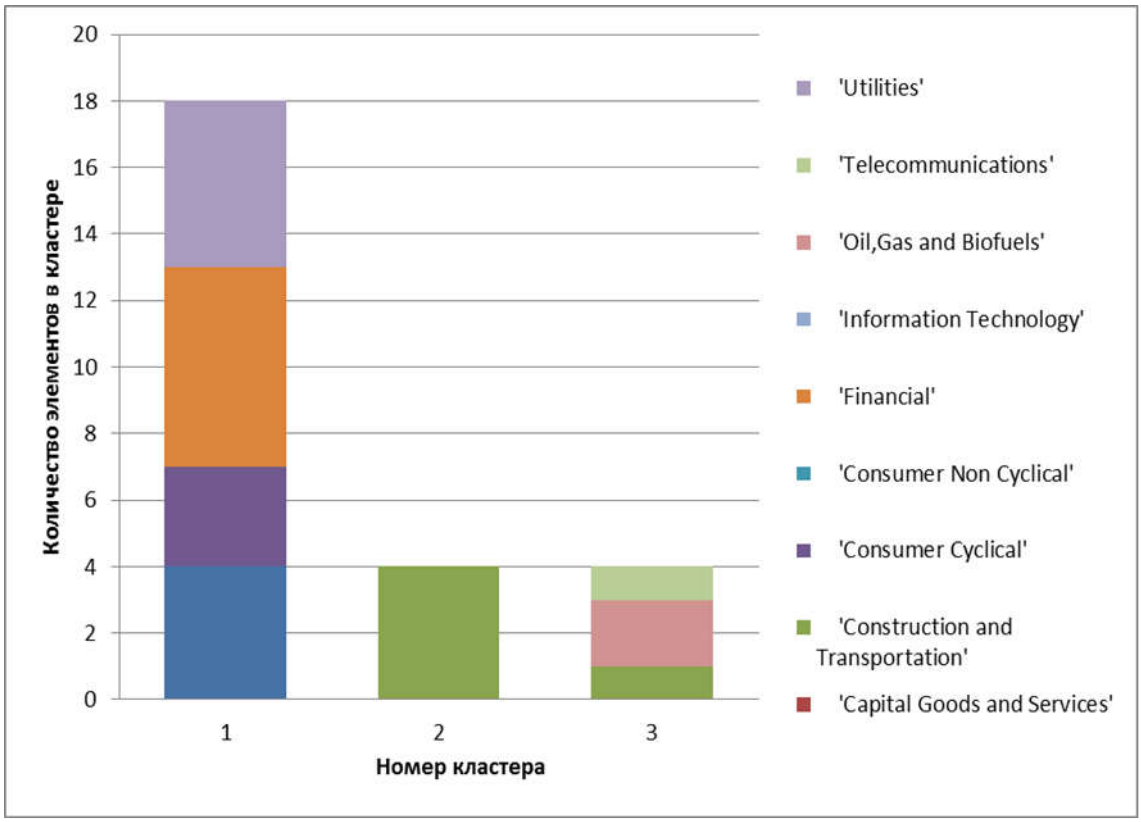

Рис. 2. Состав кластеров, найденных модифицированным методом MST.

Заметим, что во второй кластер попали исключительно активы из сектора *«Construction and Transportation»*. При этом третий кластер образуют элементы из *«Oil,Gas and Biofuels»* и *«Telecommunication»*. Важно отметить, что эти два сектора не присутствуют ни в одном другом кластере. Что касается первого кластера, то он получится достаточно большим и содержит в себе представителей секторов, не вошедших в другие кластеры. Поэтому, разумно, попытаться разделить его еще на четыре кластера, так как именно такое количество секторов представлено в нем. Результаты изображены на Рисунке 3, исходя их которого, можно сделать два замечания:

1. Метод кластеризации на основе MST постоянно выделяет одну большую компоненту, в состав которой входят представители разных секторов экономики;

2. Третий кластер образован только из элементов *«Consumer and Cyclical»,* не вошедших в другие кластеры.

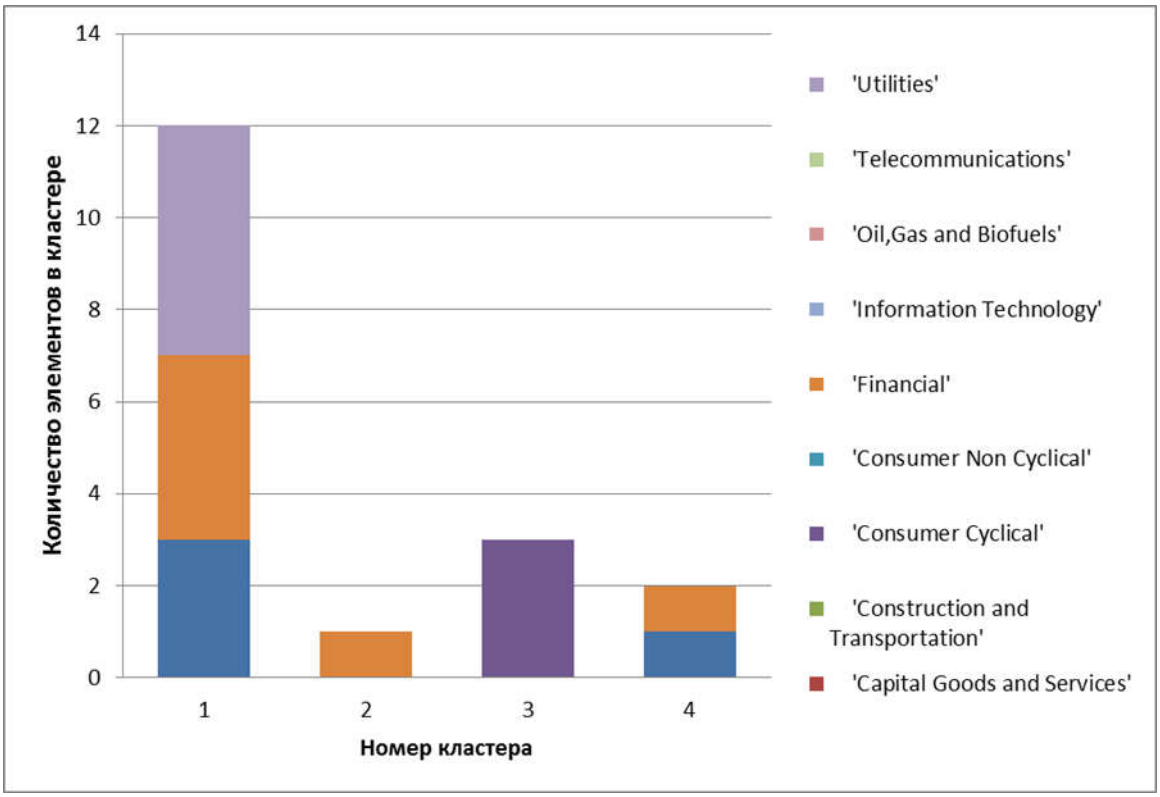

Рис. 3. Состав кластеров, найденных модифицированным методом MST.

При дальнейшем разбиении наибольшей компоненты не удалось извлечь содержательной информации, поскольку в результате наблюдается преобладание большого числа единичных кластеров. В целом, попытку применения кластерного анализа на основе MTS можно считать удачной, так как получилось выявить группы активов входящих в различные следующие сектора экономики: *«Consumer and Cyclical», «Construction and Transportation», «Oil,Gas and Biofuels»* и *«Telecommunication».*

Говоря о результатах, полученных применением других методов - можно утверждать о явном отсутствии содержательной информации кластерных структур с точки зрения индустриальной классификации активов. В случае метода кластеризации на основе метода максимизации функции модулярности этот факт может быть вызван тем, что модулярность больше нацелена на выявление сообществ в социальных сетях, где не требуется фильтрация графа и часто не учитывается вес ребер.

Выводы

В работе рассмотрена задача кластеризации в применении к выделению связанных структур (кластеров) на финансовых рынках. Для анализа качества разбиения рассмотрены индексы оценки качества, такие как кластерно-векторный баланс, индекс Дэвиса-Болдина и модулярность. Установлено, что лучшими, с точки зрения качества разбиения, являются Евклидова метрика, а также динамическая трансформация временной шкалы (DTW). Несмотря на высокую вычислительную сложность $O(n^2)$, *DTW* учитывает специфику временных рядов и значительно улучшает качество разбиения на основе метода К-средних для финансовых рядов.

В ходе анализа метода кластеризации на основе минимального остовного дерева (*MST)* было установлено, что метод выделяет несколько кластеров единичного размера и одну большую компоненту. В работе предложена модификация этого метода для получения более сбалансированных кластеров. Новый подход позволил получить содержательную интерпретацию найденных кластеров. При интерпретации кластерных структур, полученных другими методами, содержательной информации обнаружено не было. Это может свидетельствовать о том, что исследуемые метрики и методы в меньшей мере приспособлены к извлечению неявных знаний о внутренних закономерностях и структуре данных с финансовых рынков.

Литература

1. Arthur, D. S. Vassilvitskii S. How slow is the k-means method? // ACM New York, NY, USA. -2006. –Pp. 144–153.

2. Brandes, U. D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, D. Wagner On modularity clustering // IEEE Transactions on Knowledge and Data Engineering. – 2008. – No. 2. – Pp. 172-188.

3. Clauset, A. Newman M. E. J. . Moore C., Finding community structure in very large networks // Physical Review E 70, 066111. – 2004.

4. Davies, D.L. Bouldin D.W. A cluster separation measure // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1979. – No. 2. – Pp. 224–227.

5. Djauhari, M., & Gan, S. Optimality problem of network topology in stocks market analysis. // Statistical Mechanics and Its Applications, -2015. -No.419. -Pp. 108-114.

6. Fortunato, S. Barthélemy M. Resolution limit in community detection // Proc. National Academy, USA. -2007. –No. 104. –Pp. 36.

7. Iglesias, F. Kastner W. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns // Energies. -2013. –No. 6. – Pp. 579-597.

8. Good, B. H. de Montjoye Y. A., Clauset A. Performance of modularity maximization in practical contexts // Physical Review E 81, 046106. – 2010.

9. Liao, T.W. Clustering of time series data—A survey // Pattern Recognition. – 2005. – No. 38. – Pp. 1857–1874.

10. Newman, M. E. J. Girvan M. Finding and evaluating community structure in networks // Physical Review E 69, 026113. – 2004.

11. Rendón, E. Abundez I., Arizmendi A., Quiroz E. Internal versus External cluster validation indexes // International journal of computers and communications. -2011. –Vol. 5. –Issue. 1.

12. Song, W-M. Di Matteo T, Aste T Hierarchical Information Clustering by Means of Topologically Embedded Graphs // PLoS ONE 7(3): e31929.doi:10.1371/journal.pone.0031929. -2012.

13. Tumminello, M. Aste T., Di Matteo T., Mantegna R. N. A tool for filtering information in complex systems / // Proceedings of the National Academy of Sciences USA, -2005. -Vol. 102. -Pp. 10421-10426.