УДК 338.2

А. Н. Кисляков, Н. Е. Тихонюк

Российская академия народного хозяйства и государственной службы (Владимирский филиал), Владимир, e-mail: tasha-ti@yandex.ru

ВЫБОР МЕТОДА СЕГМЕНТИРОВАНИЯ КЛИЕНТСКОЙ БАЗЫ В УСЛОВИЯХ ИНФОРМАЦИОННОЙ АСИММЕТРИИ

Ключевые слова: сегментация рынка, методы кластеризации, информационная асимметрия.

Предметом исследования являются подходы к кластеризации потребителей товаров и услуг, относительно их поведенческой активности, а также разработка методики кластеризации клиентов на основе результатов их поведенческой активности. Важной особенностью методики является возможность работать не только с числовыми, но и с категориальными признаками, а также возможность определения количества кластеров. В результате исследования решены следующие задачи: рассмотрены различные методики классификации и кластеризации клиентов по признакам поведения и разъяснены основные категории; обоснована необходимость кластеризации клиентов на основе не только числовых, но и категориальных переменных, что позволяет при необходимости перейти от методов контролируемой классификации (например, RFM-анализа) к методам иерархической кластеризации; исследованы и предложены к использованию метрики расстояний между элементами внутри кластера, а также показатели объединения элементов в кластер и разделения отдельных кластеров между собой; исследованы методы иерархической кластеризации с использованием групп с нарушенной симметрией, обоснован выбор методик иерархической кластеризации для различных случаев; решается проблема автоматизированной оценки количества кластеров; предложена методика кластеризации на основе числовых и категориальных признаков с использованием иерархических методов разбиения на кластеры. В качестве основной причины, порождающей изменения поведенческой активности клиентов в работе указывается информационная асимметрия, которая выражается в разной степени информированности групп продавцов и групп покупателей-пользователей продукта о состоянии рынка. В результате разбиения потребителей продуктов на кластеры существует возможность установить причины нарушения симметрии в поведении групп потребителей товаров и услуг.

A. N. Kislyakov, N. E. Tikhonyuk

Russian Academy of National Economy and Public Administration under the President of the Russian Federation (Vladimir branch), Vladimir, e-mail: tasha-ti@yandex.ru

CHOICE OF A METHOD FOR SEGMENTING THE CLIENT BASE IN THE CONDITIONS OF THE INFORMATION ASYMMETRY

Keywords: market segmentation, clustering methods, information asymmetry.

The subject of the research is the approaches to clustering goods and services in relation to behavioral activity, as well as the development of methods for clustering customers based on the results of their behavioral activity. An important feature of the technique is the ability to work not only with numeric and categorical features, as well as the ability to determine the number of clusters. As a result of the study, the following tasks were solved: various methods of classification and clustering of clients according to the characteristics of behavior are considered and the main categories are explained; substantiated the need for clustering customers based not only on numerical, but also categorical variables, which makes it possible, if necessary, to switch from methods of controlled classification (for example, RFM analysis) to methods of hierarchical clustering; investigated and proposed for use metrics of distances between elements within a cluster, as well as indicators of combining elements into a cluster and separating individual clusters among themselvesthe methods of hierarchical clustering with the use of groups with broken symmetry are investigated, the choice of methods of hierarchical clustering for various cases is justified; the problem of automated estimation of the number of clusters is being solved; a clustering technique based on numerical and categorical features using hierarchical clustering methods is proposed. Information asymmetry is indicated as the main reason that generates changes in the behavioral activity of customers in the work, which is expressed in different degrees of awareness of the groups of sellers and groups of buyers-users of the product about the state of the market. As a result of dividing consumers of products into clusters, it is possible to establish the reasons for the violation of symmetry in the behavior of groups of consumers of goods and services.

Введение

В условиях цифровизации экономики особую степень важности приобретает персонализированный подход к взаимодействию с клиентами, которые являются потребителями товаров и услуг. В целях определения намерений потребителей удобнее всего разбить их на группы-сегменты, объединенные по различным признакам, выявленным по мере анализа потребностей, то есть выполнить кластеризацию.

Цель работы – разработка методики сегментирования клиентов на основе результатов их поведенческой активности с использованием категориальных признаков.

Предлагаемый подход к разработке моделей кластеризации может быть использован обработки данных, которые обладают различной степенью детализации, а также позволяет найти метод кластеризации клиентского портфеля компании в условиях информационной асимметрии.

Информация о поведении потребителей необходима для выбора между наборами предопределенных действий, которые необходимо предпринять. Этот тезис особенно важен при описании поведенческих особенностей потребительских сегментов, или кластеров потребителей. Особую ценность этот метод получил при внедрении цифровых технологий работы компании в условиях массового перехода к цифровому маркетингу. Простейшее использование информации в процессе принятия решений – это единичное условие «если-то» (если цена ниже некоторой суммы, покупайте; если она выше этой суммы, продавайте) [10]. В большинстве исследований информация используется главным образом для определения правил процессов принятия решений (например, для определения цены, по которой человек будет покупать или продавать, или для установления набора правил, которым будут следовать отдельные лица), в которых основное внимание уделяется принятые решения и последствия этих решений.

При этом всегда первым шагом является формирование определенных групп потребителей, обладающих схожими характеристиками. В маркетинге это называется сегментацией потребителей и является первым шагом любой маркетинговой практики.

Проведение сегментации клиентов необходимо в том числе для того, чтобы изучить сегменты на предмет их свойств и различных способов применения маркетинговой тактики к этой конкретной группе. Необходимо сравнить конкурирующие брендов и исследования поведения различных сегментов потребителей по отношению к ним. Следовательно, модель сегментации сможет эффективно повысить прибыльность и конкурентоспособность компании.

Материалы и методы исследования

Существует целый ряд достаточно эффективных методик разделения объектов по признакам (кластеризации объектов), как универсальных [5], так и специализированных [4], которые ориентированы на исследование результатов поведенческой активности покупателей.

При кластеризации объектов набор данных разделен на несколько групп, и точки данных в каждой группе, то есть внутри кластера больше связаны друг с другом, чем с теми, что находятся в других кластерах. Эти точки данных объединяются путем обнаружения соответствий в соответствии с признаками, обнаруженными в необработанных данных, однако основная цель этого анализа- найти подходящее количество кластеров, которые являются релевантными, а также полезными для целей анализа. Этот процесс представляет собой повторяемую и итеративную задачу, при которой огромные объемы необработанных данных сканируются на предмет сходства и закономерностей.

Клиенты различаются по поведению, потребностям, желаниям и характеристикам, и основная цель методов кластеризации – идентифицировать разные типы клиентов и сегментировать клиентскую базу на кластеры схожих профилей, чтобы процесс целевого маркетинга мог выполняться более эффективно. Как иерархические, так и неиерархические алгоритмы кластеризации широко используются в сегментировании клиентов.

Эвристические и экспертные методы сегментации (маркетинговый термин) или кластеризации (термин теории систем) на основе контролируемых признаков, подобно ABC-XYZ анализу, RFM-анализу достаточно просты, но весьма относительны. Недостатки их состоят в том, что используется всего лишь несколько (не более 2-3) признаков поведения клиентов для описания их поведения, исключая из рассмотрения прочие факторы: стабильность рынка, например. Количество групп клиентов определяется заранее. Например, в классическом RFM-анализе используется 9 групп клиентов. Кроме этого, к недостаткам проведения кластеризации по клас-

сическим методам можно отнести то, что она проводится на определенную дату, и не отражает особенности поведения потребителя.

Методы неконтролируемой кластеризации необходимы для нахождения сегментов среди клиентов несут в себе основную идею разделить клиентов на группы без использования предварительных гипотез о характеристиках каждой группы и исходя из имеющихся данных.

Чаще всего на практике используется набор методов (K-mean, C-mean, иерархическая кластеризация и т.п.), которые позволяют определить близость объектов (метрическое расстояние) друг от друга на основании их свойств. Это необходимо чтобы сгруппировать объекты (сегменты потребителей) для целей эффективной реализации маркетинговой политики. Клиент описывается вектором (набором) признаков, каждый элемент этого вектора описывает какую-то характеристику клиента (покупка товара определенной категории, количество дней с момента последней покупки, и т.п.). После чего этот вектор преобразуется в определенный формат, и на основе подсчета метрик расстояний между этими векторами на выходе получается разделение клиентов на кластеры.

Проведенный анализ указанных методик позволяет выявить основные проблемы кластеризации покупателей:

1. Наиболее простые методы (такие как K-mean, C-mean) требуют первоначального определения количества групп-кластеров.

2. Необходимость выбора метрики расстояний между элементами внутри кластера и между центрами всех кластеров в целях дальнейшей адекватной интерпретации результатов.

3. Признаки сделок являются факторами, которые могут коррелироваться между собой, эти особенности необходимо учитывать при интерпретации результатов.

Все это не позволяет использовать универсальные модели для выделения сегментов рынка.

Наибольшую сложность имеют два вопроса. Во-первых, это выбор метрик расстояний между элементами пространства признаков, по которым выполняется кластеризация клиентской базы.

«Исходными данными для проведения кластерного анализа служит матрица расстояний между объектами, сформированная с использованием той или иной метрики. Распространенная мера удаленности объектов друг от друга, используемая чаще всего – евклидово расстояние [1].

Однако, в случае сильно разреженных данных (например, маркетинговые активности в компании проводятся не регулярно, или спрос на товар подвержен серьезным колебаниям), следует обращать внимание на заказы, а не на их отсутствие. В этом случае кластеризации факты совершения сделки важнее чем факты отсутствия этой сделки, потому как факт отсутствия сделки не означает что клиента не интересует данный товар. И становится актуальным сравнение качественных признаков сегментов для лучшего понимания причин маркетинговой активности. В этих случаях стандартные методы не могут быть использованы, поэтому нужен расчёт расстояния, например, расстояния по косинусу [1].

Метод неконтролируемой кластеризации используется в случае, когда задание метрики расстояния между фактами позволит более точно разбить пространство признаков в случае сильно разреженных данных.

После расчета матрицы расстояний между объектами необходимо последовательно объединить объекты в кластеры. В основе метода, который позволяет реализовать более точное разбиение на кластеры, лежит теория графов и метод дендрограмм с переходом к категориальным данным [1].

Для разбиения на кластеры также необходимо определить метод объединения в кластеры, т.е. метод который позволит выявить наиболее сильные связи между группами объектов.

Заключительным этапом кластерного анализа является выбор максимального количества кластеров, на которые следует разделить совокупность объектов. Для количественного обоснования оптимального количества необходимо использовать подход на основе анализа силуэтов. Ширина силуэта i-го объекта (s_i) определяется соотношением

$$s_i = \frac{b_i - a_i}{\max\left(b_i, a_i\right)}, \qquad (1)$$

где a_i – среднее расстояние между i-м объектом и всеми членами кластера, которому он принадлежит (если объект в кластере один, то $a_i = 0$),

b_i – среднее расстояние между i-м объектом и членами другого ближайшего кластера (на практике рассчитывается среднее расстояние до членов всех остальных кластеров и выбирается минимум).

Для всех объектов классификации рассчитывают ширину силуэта, чем выше этот показатель, тем надежнее классификация. Затем вычисляют среднюю ширину силуэта для количества кластеров от 2 до *M*. Оптимальной считается классификация с наименьшей средней шириной силуэта [8].

Во-вторых, необходимо учесть тот факт, что реакция покупателей или их поведенческая активность могут зависеть от множества случайных факторов, нарушая баланс интересов участников рыночных отношений. Это явление называется информационной асимметрией [2] и выражается в разной степени информированности групп продавцов и групп покупателей-пользователей продукта о состоянии рынка, что определяет различные поведенческие настроения и намерения участников рынка. Это особенно актуально при работе с промышленным сегментом, где осуществляются разовые закупки с большим интервалом.

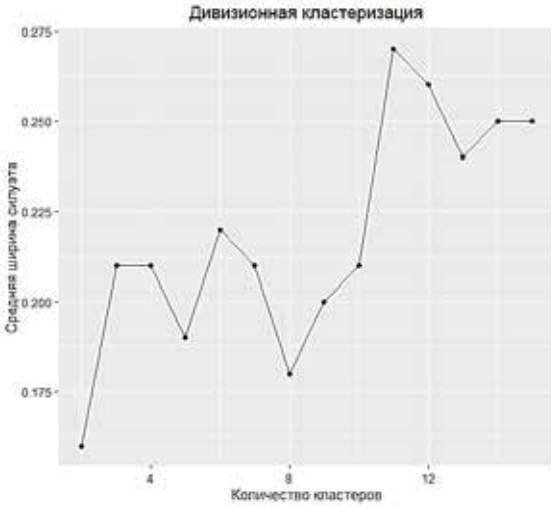
Проведем анализ данных потребительского портфеля предприятия машиностроительного сектора.

В результате RFM анализа (метод контролируемой классификации) были разделены на группы, исходя из данных о давности проведения заказа, частоты (количества сделок) и величины оказанной услуги (средний чек по каждой организации).

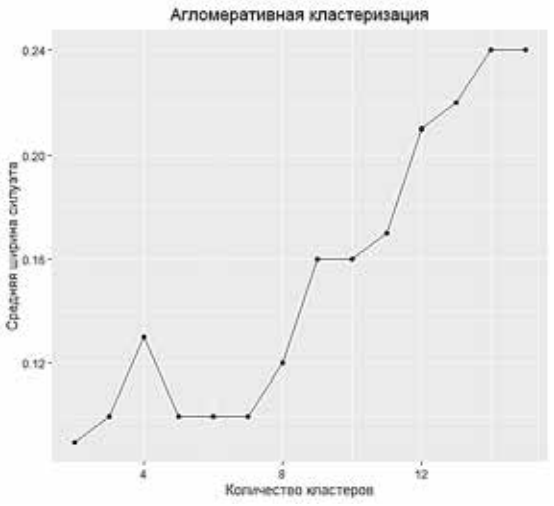
В результате анализа было выявлено 7 групп клиентов. Но этот вариант иерархической кластеризации не позволил ответить на вопрос: а отличается ли поведение клиентов стабильным трендом? Или это случайное значение на конец периода?

Для ответа на этот вопрос проведен анализ клиентского портфеля методам иерархической кластеризации.

Были получены следующие значения, приведенные на рисунке 1. Резкое изменение показателя суммы расстояний между объектами внутри кластера (резкий изгиб на графике) при определенном количестве кластеров позволяет сделать вывод о том, что дальнейшее разбиение на кластеры теряет смысл. В данном случае (рисунок 2) сумма квадратов расстояний между объектами внутри кластера не дает однозначных выводов о выборе количества сегментов. Однако в случае агломеративной кластеризации изгиб графика более плавный. И в том и в другом случае логично предположить, что в данной выборке присутствует 5-6 кластеров.



a



b

Рис. 1. Средняя ширина силуэта для дивизионной (a) и агломеративной (b) кластеризации

Иным образом обстоят дела с показателем ширины силуэта. При использовании метода оценки силуэтов, следует выбирать такое количество, которое дает максимальную ширину силуэта, потому что вам нужны кластеры, которые достаточно далеко отстоят от друга, чтобы считаться отдельными. Если набор данных будет разбиваться на более мелкие группы, тем лучше кластеры отделимы друг от друга, однако так дело может дойти до отдельных точек и процесс потеряет смысл, поэтому целесообразнее оценивать локальные экстремумы [1]. Как видно из рисунка 5 ширина силуэтов в обоих случаях оптимальна при 4-5 кластерах.

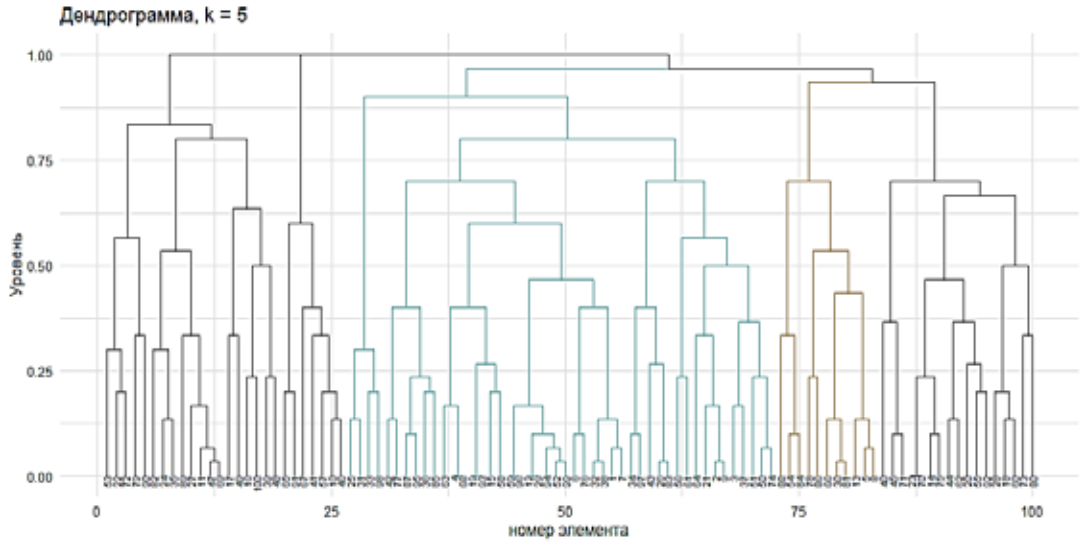


Рис. 2. Дендрограмма на основе агломеративной кластеризации при количестве кластеров равном 5 (составлено авторами)

При прочих равных показателях агломеративная кластеризация в данном случае является более сбалансированной относительно количества объектов в кластере и дендрограмма (рисунок 2) выглядит более симметрично. Симметричность кластеров важна по нескольким причинам. Во-первых, это дает нужную информацию о стабильности поведения потребителей внутри группы. Это означает, что можно планировать реакцию на маркетинговые стимулы с увеличенной точностью. Во-вторых, симметричность является также признаком устойчивости системы.

Таким образом, уточнение количества кластеров (сегментов) на рынке приводит к большей концентрации маркетинга на целевых потребителях, и как следствие более эффективному маркетингу. На основе структуры клиентского портфеля можно сделать вывод, что в рамках работы с клиентами и позиционирования компании в конкурентной среде стоит применять новые маркетинговые инструменты, причем разные для отдельных сегментов. Внедрение цифрового маркетинга позволит обрести компании конкурентное преимущество, заключающееся в более современных методах коммуникации с клиентами, например, позволяющие напомнить о сотрудничестве одной группой клиентов или перевести компании из другого сегмента в более высокую категорию А с помощью e-mail рассылки персональных предложений, а также увеличить количество точек контакта с клиентами, создавая веб-представительства компании в социальных сетях или на электронных торговых площадках.

Заключение

Из-за внедрения новых цифровых технологий данные о потребителях растут в геометрической прогрессии. При работе с таким большим объемом данных организациям необходимо использовать более эффективные алгоритмы кластеризации для сегментации клиентов. Эти модели кластеризации должны обладать способностью эффективно обрабатывать эти огромные данные. Каждый из рассмотренных выше алгоритмов кластеризации имеет свой набор достоинств и недостатков. Вычислительная скорость алгоритма кластеризации K-средних относительно лучше по сравнению с алгоритмами иерархической кластеризации, поскольку последние требуют вычисления полной матрицы близости после каждой итерации.

Кластеризация K-средних дает лучшую производительность для большого количества наблюдений, в то время как иерархическая кластеризация имеет возможность обрабатывать меньшее количество точек данных. Основным препятствием является выбор числа кластеров «K» для процесса K-средних, который должны быть предо-

ставлены в качестве входных данных для этого алгоритма неиерархической кластеризации. Это ограничение не существует в случае иерархической кластеризации, поскольку для нее не требуются какие-либо центры кластеров в качестве входных данных. Выбор групп кластера, а также их количество зависит от пользователя.

Иерархическая кластеризация также дает лучшие результаты по сравнению с K-средними при использовании случайного набора данных. Выходные данные или результаты, полученные при использовании иерархической кластеризации, представлены в форме дендрограмм, но выходные данные K-средних состоят из кластеров с плоской структурой, которые может быть трудно анализировать. Таким образом, алгоритмы разделения, такие как K-Means, подходят для больших наборов данных, в то время как алгоритмы иерархической кластеризации больше подходят для небольших наборов данных.

Выводы

Анализ результатов позволяет сделать следующие выводы:

1. Для получения более сбалансированных результатов и осмысленной интерпретации результатов кластеризации необходим переход от числовых переменных к категориальным.

2. Косинусное расстояние используется для сильно разреженных данных при различной важности фактов поведенческой активности клиентов.

3. Использование расстояния Говера в качестве метрики кластеризации позволяет выполнить расчет расстояний между числовыми и категориальными переменными.

4. Использование категориальных данных делает невозможным применение известных методов кластеризации, однако позволяет выполнить более «плавный» переход от методов контролируемой классификации (например, RFM-анализа) к методам иерархической кластеризации.

5. Методы иерархической кластеризации отлично подходят для оценки необходимого количества кластеров и дают возможность автоматизированной оценки количества кластеров.

6. Агломеративная кластеризация на основе использования дендрограмм дает более точные результаты относительно количества объектов в кластере, что позволяет судить о более однородных характеристиках объектов внутри кластера.

Методика, предлагаемая авторами, может быть использована в следующих случаях:

1) При проведении сегментации клиентской базы для проведения таргетированных маркетинговых мероприятий;

2) Выделение наиболее типичных представителей кластеров клиентов для формирования точечного ассортиментного предложения;

3) Формирование спроса на продукт в зависимости от типа клиента;

4) Повышение конверсии в покупателей за счет понимания алгоритмов принятия решения о покупке.

Для использования в бизнесе визуализация данных составляет основную часть эффективного анализа данных, а иерархическая кластеризация помогает в этом. С учетом недостатков и достоинств этих двух методов выясняется, что сочетание лучших из этих алгоритмов может превзойти отдельные модели. Таким образом, можно использовать разные алгоритмы кластеризации из-за их свойств по отношению к разным типам данных. последовательно так, чтобы можно было полностью использовать преимущества этих методов. Однако процесс выбора этих подходящих методов, а также их разумное внедрение может потребовать значительных временных затрат на изучение и обработку данных наряду с адекватным пониманием целей и требований.

Вместе с тем, вопрос асимметричности в структуре клиентского портфеля остается пока не изученным. Появление «случайных» отклонений при формировании клиентского портфеля может привезти к существенному долгосрочному росту, если это отклонение связано с изменением трендов. Это может быть в дальнейшем темой отдельной работы.

Исследование выполнено в рамках работ по гранту РФФИ 18-07-00170 А «Создание прогностических моделей эволюции природных, живых и социально-экономических систем на основе конечных групп нарушенной симметрии».

Библиографический список

1. Кисляков А. Н., Поляков С. В. Иерархические методы кластеризации в задаче поиска аномальных наблюдений на основе групп с нарушенной симметрией // Управленческое консультирование. 2020. № 5. С. 116–127.

2. Тихонюк Н.Е., Кисляков А.Н. Экономические модели работы с асимметрией информации: эволюция подходов // Региональная экономика: опыт и проблемы. Материалы XI международной научно-практической конференции (Гутманские чтения) 15 мая 2018 года / под общ. ред. А.И. Новикова и А.Е. Илларионова. Владимир: Владимирский филиал РАНХиГС, 2018. С. 236-244.

3. Якимов В.Н., Шурганова Г.В., Черепенников В.В., Кудрин И.А., Ильин М.Ю. Методы сравнительной оценки результатов кластерного анализа структуры гидробиоценозов (на примере зоопланктона реки Линда Нижегородской области) // Биология внутренних вод. 2016. № 2. С. 94-103.

4. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning with Applications in R // Publisher: Springer. 2013.

5. Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition // Publisher: Springer. 2017.

6. Alboukadel Kassambara. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis). 2017.

7. Газиев Г.З., Курдюкова Г.Н., Курдюков В.В. Кластеризация Big Data для их анализа и обработки // Современный взгляд на будущее науки: приоритетные направления и инструменты развития: сборник научных статей по итогам международной научно-практической конференции. СПб.: Редакционно-издательский центр «КУЛЬТ-ИНФОРМ-ПРЕСС», 2017.

8. Печеный Е.А. Динамическая кластеризация потока больших данных, 2019.

9. Демидова Л.А., Степанов М.А. Подход к решению задачи выявления структурных трансформаций в группах временных рядов. Cloud of Science, 2019.

10. Bao L., Fritchman J.C. Information of Complex Systems and Applications in Agent Based Modeling. Sci. Rep., 2018.

11. Laureti P., Zhang Y.-C. Matching games with partial information. Phys. A: Stat. Mech., 2003.

12. Wang Y., Li Y., Liu M. Impact of asymmetric information on market evolution, 2007. 665 p.