ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ЕКОНОМІЦІ

УДК 519.2

Ю.І. Мінаєва

Київський національний університет будівництва і архітектури, Київ

СТРАТЕГІЯ ВИБОРУ ПОРТФЕЛЯ ЦІННИХ ПАПЕРІВ, ПОДІБНОГО ДО РИНКУ ЦІННИХ ПАПЕРІВ ЗА ПОВОДЖЕННЯМ

Розглянуто стратегію формування портфеля цінних паперів на підставі методів ієрархічної кластеризації. Вказано на можливість використання для розв'язання задачі дендрограм. Зазначенощо сполучення кластерного аналізу з р-адичним аналізом дозволяє істотно спростити проблему математичного аналізу бінарних дерев (результат ієрархічної кластеризації). Наведено приклад.

Ключові слова: портфель цінних паперів, цінні папери, кластерний аналіз, дендрограма

Вступ

Дослідження ринку цінних паперів (ЦП), як тільки економіка сформувалася як наука, представляло й представляє винятково актуальну й складну задачу [1]. Як правило, розв'язання цієї проблеми зведено до розв'язання сукупності комплексів задач:

– одержання максимально можливої інформації про ринок;

– обробка (статистична, аналітична й ін.) отриманої інформації й формування нових ознак, що характеризують ринок;

– ухвалення рішення про поведінку інвестора (купувати або продавати ЦП відразу або з деякою часовою затримкою).

Робота на ринку до останнього часу базувалася на тих засадах, що фінансові процеси, які протікають на ринку, зміна цін цінних паперів підкоряються статистичним законам, і тому зазвичай трейдери намагались отримати прибуток , не враховуючи або недостатньо враховуючи поводження фінансового ринку, його хаотичну природу.

Саме тому в даний час актуальними є методи, які дозволять ефективно розв'язати задачу формування портфеля за умов невизначеності і дозволять врахувати хаотичну природу ринку. До таких належать методи, засновані на застосуванні нових (зокрема, фрактальних) принципів для моделювання поводження ринку. Найбільш повно такий підхід показаний у роботах [2-5]. Нижче проводиться короткий огляд роботи Б.Вільямса, де антитеза «мічурінської» ідеології проявляється найбільш яскраво.

Як відзначає Б. Вільямс, «одержання прибутку досягається не протистоянням ринку: потрібно просто танцювати в такт із ним. Ринки – природне явище і їх діяльність не підкоряється законам класичної фізики, параметричної статистики або лінійної математики». Виходячи з цього, підхід до обробки інформації, що найбільш повно відповідає природі ринку , варто шукати на рівні фрактальної теорії.

За думкою Б.Вільямса, сьогодні необхідне повернення до здорового глузду й правильного погляду на торгівлю (цінними паперами), згідно з яким необхідно враховувати два види знань:

1) основної структури ринку;

2) власних базових принципів учасника ринку, роботи на ринку.

Прийняття теорії хаосу як головної парадигми функціонування ринку припускає облік непередбачуваності ринку, отже, необхідно підійти до торгівлі за допомогою методів, які не ґрунтуються на прогнозах. Б.Вільямс не без підстав затверджує, що всі ми працюємо на ринку, використовуючи системи *наших* (суб'єктивних) уявлень, не маючи у своєму розпорядженні знання про внутрішню структуру ринку. Тільки тоді, коли ми розкриємо *основну структуру ринку* й приведемо у відповідність наше бачення процесів, що відбуваються в дійсності, то отримаємо результат – успіх на ринку.

В роботах [2-4] відзначається, що видатні вчені Пригожин, Файгенбаум, Бернсли, Смейл і Хенон та інші виявили, що на межі між конфліктами протилежних сил (а ринок – це в загальному випадку середовище з конфліктними інтересами) знаходиться не народження хаотичних, безладних структур, як вважалося раніше, а відбувається спонтанне виникнення самоорганізації порядку більш високого рівня. Більше того, структура цієї самоорганізації не структурована відповідно до схем Евкліда, Ньютона, а є новим видом організації. Автор [1] стверджує, що «зважаючи на все, організація цього порядку застосовна до всього: від застібок блискавки до економічного ринку».

Б. Мандельброт [2] виявив близьке споріднення між *фрактальним числом* ріки Міссісіпі й цінами на бавовну на всьому часовому інтервалі, що він вивчав. Значення цього спостереження неможливо недооцінити. Воно означає, що ринки є "живою" нелінійною функцією. Це частково пояснює, чому 95 відсотків трейдерів, що використовують звичайний технічний аналіз, постійно програють. Мало того, що технічний аналіз заснований на помилковому припущенні про подобу майбутнього минулому, в ньому часто використовуються невідповідні лінійні методи досліджень.

Відзначимо, що запропонований у роботі спосіб оцифровки дендрограм дозволяє одержати аналог фрактального числа ріки Миссісіпі у вигляді величини $\mathbf{C}_x$ або значень максимальних рангів для ринку й для портфеля, він також показує, що одержати *повний збіг* цих величин теоретично неможливо, але поводженням на ринку (покупка/продаж ЦП) можна домогтися досить близького їх збігу.

Відзначено [1], що якщо трейдер бажає змінювати плин своєї торгівлі, він повинен працювати зі *своєю основною структурою*, а не з поводженням, відтвореним нею. Як тільки структура ляже в основу торгівлі, трейдер одержить новий поштовх, що, подібно потоку ріки, створює необхідні імпульси й допомагає досягти бажаних результатів.

Основна концепція, що сформульована фахівцями ринку, полягає в тому що [25]: на початковому етапі необхідно визначити основну структуру, яка буде вести діяльність на ринку, а потім змінювати її (діяльність) так, щоб можливо було створювати те, що дійсно потрібно отримати від ринку, результати в економіці і житті.

На підтвердження сформульованих принципів можемо зробити наступний висновок: структура визначає поводження, структура визначає поведінкову модель будь-чого: від урагану до ринку, у тому числі структура торговельних площадок визначає поводження трейдерів в торговельній ямі. В [1] наведена заповідь дзен-буддизму: «Спостерігайте природний порядок речей. Шукайте відповідність цьому порядку, а не протидійте йому, щоб він не протидіяв вам».

Звернемо увагу на наступну категоричну рекомендацію: «*Перше, що ... повинні зрозуміти й прий- няти: ми ніколи не зможемо вплинути на ринок. Єдино вірне рішення - прямувати за ринком і бути впевненими, що ми дійсно прямуємо за ринковим рухом*».

Одним із способів визначити структуру об'єктів є побудова бінарних дерев (дендрограм) на основі методів і моделей кластерного аналізу. Останнім часом кластерний аналіз переживає буквальний ренесанс, пов'язаний з використанням ультраметрики й p-адичного аналізу стосовно до нових класів задач. Застосування p-адичної метрики представляє по суті альтернативний підхід до розв’язання традиційних задач кластеризації, дослідник при цьому не тільки зіштовхується з множиною складностей, тому що застосовуючи p-адичну метрику і відповідно неархімедову геометрію, ми втрачаємо можливості застосування в чистому вигляді всієї теорії класичної геометрії, але одержує сприятливі можливості до постановки й розв’язання нових задач [5-10].

Кластерний аналіз у p-адичному базисі, за визначенням одного з відомих фахівців у цій області Муртага (Fionn Murtagh), дозволяє з висоти пташиного польоту побачити проблему [5], додамо, природно, втративши певні деталі. Кластерному аналізу в p-адичному базисі присвячена величезна кількість публікацій, хоча на російській або українській мовах ця проблема практично не освітлена, особливо її нові додатки. Таким новим додатком, на наш погляд, є проблема аналізу s вибору портфелів ЦП в p-адичному базисі, розв’язанню якої присвячена дана робота.

Стан проблеми

Кластер-аналіз як спосіб угруповання багатовимірних об'єктів, заснований на поданні результатів окремих спостережень точками придатного геометричного простору з наступним виділенням груп як «згустків» цих точок, потенційно найбільш раціональний під час розв’язання задач, які можна віднести до групи таких, де тією чи іншою мірою потрібна «*подібність*», наприклад виділення груп підприємств, що діють у *подібних* умовах або зі *схожими* результатами, однорідних (*подібних*) груп населення за різними аспектами життєдіяльності або за способом життя у цілому. [11-15].

Однак кластерний аналіз дозволяє сформулювати новий клас задач – визначення складу портфеля ЦП, поводження якого буде *схожим* на поводження ринку ЦП у цілому, якщо під *поводженням* розуміти зростання і падіння вартості ЦП. Інакше кажучи, необхідно вирішити дві основні задачі:

– визначити наскільки існуючий портфель *схожий* на ринок;

– яким чином можна зробити портфель *схожим* на ринок.

Останнім часом є поширеною одна з рекомендацій – не проводити активну політику і додамо – чекати разом з ринком підвищення ефективності свого портфеля. За умов кризи переважна більшість інвесторів втрачає свої кошти, одиницям вдається опинитись у виграші, в той же час між ними є невелика група інвесторів, яка залишається при своїх вкладеннях, не втративши коштів, але і не вигравши.

В даний час така стратегія є досить слушною, ми можемо сказати, що в цьому випадку портфель такого інвестора деякою мірою є подібним до самого ринку. Але головне питання полягає в тому, коли саме інвестор може спокійно чекати зростання ринку.

Ми можемо сказати, що коли інвестор твердо впевнений у тому, що його портфель є певною копією ринку, яка коливається, спадає і зростає разом з ринком, він може спокійно очікувати зростання на ринку. Інакше інвестор має пересвідчитись у надійності стратегії інвестування. Для цього треба знати, що являє собою ринок сьогодні.

Запропонований метод полягає в тому, щоб визначити стан ринку цінних паперів (ЦП) і сформувати такий портфель цінних паперів (ПЦП), який би був подібним до ринку за своєю поведінкою. Для цього пропонується використання методів ієрархічної кластеризації [23].

Можливість розв’язання поставлених задач заснована на використанні *віртуальної* ієрархічної структури об'єктів, що пов'язано з відомою парадигмою Г. Саймона [16] про ієрархічну природу економічних об'єктів. Структурний підхід (кластерний аналіз і візуалізація даних) припускає виділення компактних груп об'єктів, віддалених один від одного, відшукує «природну» розбивку сукупності на області угруповань об'єктів. Цей підхід використовується для двох видів вихідних даних: матриць близькості або відстаней між об'єктами, представленими у вигляді точок у багатовимірному просторі.

Можна виділити третій напрямок розв’язання задачі кластеризації – *апроксимаційний* [14], основна ідея якого полягає в наступному: відношення, закладені у вихідних даних (ринок), потрібно якнайкраще апроксимувати іншим відношенням (портфель), що відповідає нашому уявленню про класифікацію. Класифікація звичайно задає відношення еквівалентності, при цьому вихідні дані можуть бути відбиті по-різному. Наприклад, матрицю «об'єкт-об'єкт» відстаней можна розглядати як метризовану толерантність. Тоді задача полягає в пошуку відношення еквівалентності, найближчого (у деякому сенсі) у вихідній толерантності. Можливі й інші постановки такого виду. Характер відношення, що знаходять як результат класифікації, розглянутий в роботах [14].

Найцікавіше відношення довільної структури. Показано [14], що найбільш загальний спосіб аналізу структури множини – апроксимація його деяким відношенням з довільної (заздалегідь заданою) структурою. Така задача, загалом кажучи, виходить за рамки кластер-аналізу. Але її можна розглядати більш розширено, зв'язавши довільну структуру зі способом операціоналізації деякої «супермети» класифікації.

Якщо позначити шукане відношення довільного типу через B, вихідні дані через X, а оператор переходу від X і Y – через P (не конкретизуючи вид цих конструкцій), то в загальному випадку виникає природний функціонал, що відбиває прагнення максимально наблизити результуюче відношення до наявних даних: $\|Y\text{-}PX\|$ (min, де $\|.\|$ – яка-небудь норма. Задачі такого типу – апроксимація «погано структурованої» множини X «добре влаштованою структурою» Y – відомі в математиці й мають множину додатків. Дуже часто вони є некоректними за Колмогоровим-Тихоновим.

Результати роботи всіх ієрархічних процедур кластеризації звичайно оформляються у вигляді *бінарного дерева* – дендрограми, де по горизонталі показані номери об'єктів, а по вертикалі – значення міжкласових відстаней p_{lq}, при яких відбулося об'єднання двох даних класів. Важливою перевагою ієрархічних алгоритмів є наочність результатів роботи, що дозволяє ретельно вивчити дендрограму й зробити за нею висновки, причому бажано порівнювати декілька дендрограм, що отримані різними методами.

Відзначимо наступну обставину. Розглядаючи дві підмножини об'єктів – ринок і портфель – приходимо до необхідності аналізувати ринок в «скороченому (портфельному) просторі», тобто якщо портфель, наприклад, складається з 10 об'єктів, а ринок з 1000, то він (ринок) повинен бути кластеризований до рівня портфеля, тобто мати 10 кластерів.

При скороченні простору виникає задача неспотвореного (або малоспотвореного) переносу з вихідного простору в скорочений. Звичайно намагаються зберегти геометричну структуру множини об'єктів, яка в цілому ряді випадків невідома. Можна вважати, що скорочуючи простір, орієнтуватися на всі вихідні ознаки, нераціонально, залишається або взяти за основу для порівняння вихідні властивості сукупності, або вибрати деякий зовнішній критерій скорочення розмірності й загальної обробки даних, або використовувати евристичні критерії.

Фахівці з кластерного аналізу [11-15,17] відзначають, що однією з переваг методів кластеризації є їх наочність. Всі розглянуті прийоми візуалізації базуються на одному фундаментальному положенні: людина здатна сама ухвалити рішення щодо структурі даних при їх зручному сприйнятті. Ця гіпотеза не раз перевірялася експериментально й в цілому підтвердилася. Методи візуалізації внутрішньо парадоксальні – вони використовують точні алгоритми з екстремальними властивостями лише для того, щоб згодом людина прийняла на їх основі досить наближене, природне в його розумінні рішення. Однак така парадоксальність лежить в природі економіки.

Відзначимо, що сполучення кластерного аналізу з p-адичним аналізом дозволяє істотно спростити проблему математичного аналізу бінарних дерев (результат ієрархічної кластеризації), тому що p-адична матриця дендрограми є її досить ефективною математичною моделлю.

Інтелектуальний кластерний аналіз

В останні десятиліття спостерігається зростання інтересу до нового напряму в обробці інформації – інтелектуальному аналізу даних (Data Mining - DM), що використовує методи інтелектуального кластерного аналізу [17,19]. На відміну від класичних способів аналізу, у цій галузі велика увага приділяється моделюванню поводження людини, що вирішує складні інтелектуальні задачі узагальнення, виявлення закономірностей, знаходження асоціацій і т.д. З'явився так званий *інтелектуальний кластерний аналіз*, однією з головних особливостей якого є наявність (або використання) *процедури навчання*.

Сучасна постановка задачі кластеризації розглядається в наступному вигляді. Нехай X — множина об'єктів, Y — множина номерів (імен, міток) кластерів. Задано функцію відстані між об'єктами $\rho(x, x^1)$. Є кінцева навчальна вибірка об'єктів $X^m=\{x_1, \dots ,x_m\} \subset X$... Потрібно розбити вибірку на непересічні підмножини, які називають *кластерами*, так, щоб кожний кластер складався з об'єктів, близьких за метрикою ρ, а об'єкти різних кластерів істотно відрізнялися. При цьому кожному об'єкту $x_i \in X^m$ приписується номер кластера y_i .

Алгоритм кластеризації— це функція $a\colon X \to Y$, що будь-якому об'єкту $x \in X$ ставить у відповідність номер кластера $y \in Y$. Множина Y в деяких випадках відома заздалегідь, однак частіше ставиться задача визначити оптимальне число кластерів, з погляду того або іншого *критерію якості* кластеризації. Кластеризація (навчання без вчителя) відрізняється від класифікації (навчання зі вчителем) тим, що мітки вихідних об'єктів y_i споконвічно не задані, і навіть може бути невідомою сама множина Y. Розв’язання задачі кластеризації принципово неоднозначно, тому що результат кластеризації, крім інших умов, істотно залежить від метрики, вибір якої, як правило, також суб'єктивний і визначається експертом.

Визначення способу обчислення відстані між об'єктами або групами об'єктів. Цей спосіб повинен відбивати специфіку розв'язуваної прикладної задачі. Для груп об'єктів також визначається спосіб знаходження відстані, наприклад, за принципом “віддаленого сусіда”, «ближнього сусіда » і ін. Принцип «віддаленого сусіда » найбільш виправданий в розглянутому випадку, тому що він найбільше відповідає принципу «не зберігати яйця в одному кошику».

*Актуальні проблеми кластерного аналізу***.** Задачі, поставлені в даній роботі – аналіз і формування портфеля ЦП на підставі методології ІКА – належать до важко формалізованих. Багато з цих задач характеризуються недостатністю знань про досліджувані об'єкти, що ускладнює формулювання їх математичних моделей. За цих умов одним з основних методів розв’язання задачі є *моделювання*. При моделюванні треба знати, як перерахувати результати досвіду на моделі на натуру. Для раціонального моделювання основним є поняття *подібних* явищ. Відомо, що поняття фізичної подоби природно узагальнює поняття геометричної подоби. Явища називаються подібними, якщо вони відрізняються тільки чисельними значеннями визначальних параметрів і притому так, що для них відповідні безрозмірні величини $\Pi_1, \Pi_2, \dots, \Pi_{n-k}$ збігаються.

Застосування ІКА до вибору й аналізу портфелів ЦП викликає необхідність подання результатів аналізу у формі, зрозумілій фахівцям з економіки. Як відомо, практично всі задачі кластеризації являють собою умовно екстремальні задачі. Проблема пошуку глобального екстремуму в критерії якості кластеризації – найважливіша задача в цьому напряму досліджень.

Р-Адична класифікація (кластеризація). Розглянемо питання інтелектуальної кластеризації на рівні *p*-адичного аналізу []19,20][1] .

p-adic числа. Нехай p $\in \mathbf{N}$ буде фіксованим простим числом. Тоді для будь-якого ненульового

[1] В даній роботі ІКА зводится до задачі кластеризації в 2-адичному базисі. Це пояснюється тим, що за роботами А.Хреннікова, розумова діяльність людини ефективно може бути змодельована тільки в *p*-адичному базисі [22]

$x \in q$, ми можемо завжди записати $x = p^v \cdot a/b$, для пари взаємно-простих чисел *a, b* $\in \mathbf{Z}$ та унікального $v \in \mathbf{Z}$ так, що p не ділить *ab*. В загальному випадку ціле *p*-адичне число для довільного простого *p* являє собою послідовність $x = (x_0, x_1, \ldots)$ віднімань x_n за mod p^{n+1}, що задовольняють умові $x_n = x_{n-1} (\mathrm{mod}\ p^{n+1})$, $n \geq 1$. В нотації теорії віднімань додавання і добуток цілих *p*-адичні числа визначається формулами:

$$\begin{cases}(x+y)_n = x_n + y_n (\mathrm{mod}\ p^{n+1}),\\ (xy)_n = x_n y_n (\mathrm{mod}\ p^{n+1}).\end{cases} \quad (1)$$

p-adic норма - функція $|\cdot|_p : q \to [0,\infty)$, що отримується через $|x|_p = p^{-v}$ и $|0|_p = 0$. Можна перевірити, що $|\cdot|_p$ - насправді норма на q. Крім того, $|\cdot|_p$ задоволь- няє посиленій *нерівності трикутника* (ПНТ), що полягає в тому, що для будь-яких x, y $\in$ q ми маємо $|x + y|_p \leq \max\{|x|_p, |y|_p\}$. Породжена метрика – $d_p(x, y) = |x - y|_p$ – має назву *ультраметрика*. Ультраметрика володіє цілим рядом парадоксальних властивостей, які розглядатимуться окремо.

Відносно p-adic норми q задовольняє *неархімедовим властивостям*, оскільки для кожного $x \in q$, $|nx|_p$ не буде ніколи перевищувати $|x|_p$ для будь-якого $n \in N$. Має місце повнота q щодо ультраметрики d_p, q_p – поле p-adic чисел. Більш конкретно, є унікальне подання кожного $z \in q_p$:

$$z = a_v p^v + \cdots + a_0 + a_1 p + a_2 p^2 + \cdots, \quad (2)$$

для деякого $v \in \mathbf{Z}$ і $a_i \in \{0, 1, \ldots, p - 1\}$ для всіх $i \geq v$.

Важливий підпростір q_p – одинична куля $\mathbf{Z}_p$, $\mathbf{Z}_p = \{x \in q_p \mid |x|_p \leq 1\}$, що може також бути представлене в такий спосіб:

$$\mathbf{Z}_p = \{a_0 + a_1 p + a_2 p^2 + \cdots \mid a_i \in \{0, 1, \ldots, p . 1\}, \forall i \geq 0\}. \quad (3)$$

Найважливішою властивістю p-адичних чисел є те, що вони мають ієрархічну структуру, на відміну від звичайних чисел, які розташовуються лінійно.

Цілі *p-адичні числа.* утворять кільце: їх можна додавати, віднімати й перемножувати. Однак тут відсутній природний порядок, поняття негативного й позитивного числа не мають змісту, для p-адичних чисел виконується $-1 = \lim(p^n - 1)$ при $n \to \infty$. Для величини -1 в 3-адичному базисі маємо: $-1_3 = .222222\ldots$, що відповідає дерево має вигляд (рис.1).

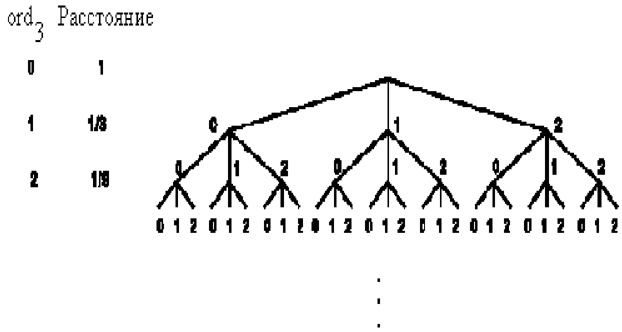


Рис.1. 3-адичне подання числа -1

P-Адична дендрограма. Як відомо, множину відстані між кожною парою даних, причому ця відстань може бути обчислена як у метричному, так і в ультраметричному базисі. Однак, якщо розмітити гілки дендрограми (наприклад, 0 (ліва гілка) та 1 (права гілка), або -1 та +1 відповідно), отриманої, скажімо, у метричному базисі, то її аналіз має виконуватися в p-*адичному* базисі. Розглянемо особливості дендрограм, користуючись результатами, викладеними в роботі [19,20].

Для аналізу даних, д*ендрограми* звичайно розмічають і ранжують (рис.2). Для ранжованої дендрограми (рис.2), створюється наступне p-*адичне* кодування термінальних вузлів, проходячи шлях від кореня: $x_1 = 0\cdot2^7 + 0\cdot2^5 + 0\cdot2^2 + 0\cdot2^1$; $x_2 = 0\cdot2^7 + 0\cdot2^5 + 0\cdot2^2 + 1\cdot2^1$; . . . $x_4 = 0\cdot2^7 + 1\cdot2^5 + 0\cdot2^4 + 0\cdot2^3$; . . .
$x_6 = 0\cdot2^7 + 1\cdot2^5 + 1\cdot2^4$ та ін.

Десяткові еквіваленти p-*адичного* представлення термінальних вузлів такі: $x_1, x_1, \ldots, x_8$ = 0, 2, 4, 32, 40, 48, 128, 192.

Відстані та норма визначені відповідно так:

$$d_p(x, x^1) = d_p\|x - x^1\| = 2^{-r+1} \text{ або } 2 \cdot 2^{-r},$$

$$\text{де } x = \sum_k a_k 2^k,$$

$x^1 = \sum_k a_k^1 2^k, r = \operatorname{argmin}\{a_k - a_k^1\}$, норма - $d_p(x, 0) = 2^{-1+1} = 1$. Для того, щоб знайти p-*адичну* відстань, розглядають найменший рівень r (якщо впорядкування йде від терміналу до кореня, рис.3), що є ідентичним парі степеневих рядів, які породжують результат 2^{-r+1}.

Отже,

$$\|x_1 - x_2\|_2 = 2^{-2+1} = 1/2;$$
$$\|x_1 - x_4\|_2 = 2^{-6+1} = 1/32;$$
$$\|x_1 - x_6\|_2 = 2^{-6+1} = = 1/32;$$

найменша p-*адична* відстань на рис.2 дорівнює 1/128. На рис. також обчислюємо найменшу p-*адичну* від-стань, яка має бути 1/128, десятковий еквівалент числа x_8 є 208.

Максимально можливий десятковий еквівалент p-*адичного* числа, що відповідає 8 термінальним вузлам – $1\cdot2^7 + 1\cdot2^6 + 1\cdot2^5 + 1\cdot2^4 + 1\cdot2^3 + 1\cdot2^2 + 1\cdot2^1 = 254$.

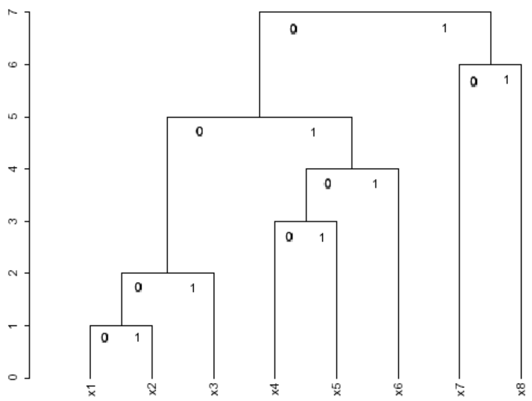
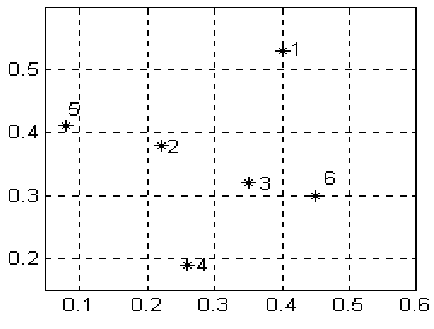


Рис.2. Розмічена та ранжована дендрограма з 8-ма термінальними вузлами, гілки помічені як 0 (ліва) та 1 (права)

Зазначимо, що р-*адичне* представлення, наведене на рис.2, не є інваріантним відносно дендрограмного представлення. Однак, якщо р-*адичні* представлення різняться для різних дендрограмних представлень, аналіз свідчить, що і р-*адична* норма і р-*адична* відстань є *інваріантними* відносно дендрограмного представлення.

При аналізі дендрограм використовують так звані агломеративні (накопичувальні) алгоритми. Розглядаючи ультраметричність, як показано у роботі [19,20], шляхом використання агломеративного алгоритму кластеризації, можемо *породити* ультраметрику (тобто отримати штучне дотриман ня ультраметричної нерівності для заданих будь-яких трьох точок) в будь-якій множині точок, забезпеченій парною функцією відмінності. Коли множина точок в просторі даних будь-якої вимірності така, що всі триплети точок задовольняють ультраметричній нерівності, то ця множина точок має *природну* ієрархічну структуру, однак не гарантується, що ця ієрархія є унікальною.

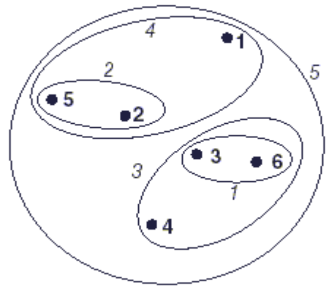
На рис.3 представлені 6 об'єктів у декартовій системі координат, обчислені метричні відстані між точками (рис. 3-а й 3-б), побудована дендрограма, у якій точки об'єднані в кластери за методом найбільш віддаленого сусіда (рис. 3-в і 3-г)



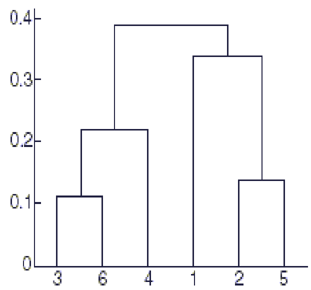
а

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0	0.24	0.22	0.37	0.34	0.23
P_2	0.24	0	0.15	0.20	0.14	0.25
P_3	0.22	0.15	0	0.15	0.28	0.11
P_4	0.37	0.20	0.15	0	0.29	0.22
P_5	0.34	0.14	0.28	0.29	0	0.39
P_6	0.23	0.25	0.11	0.22	0.39	0

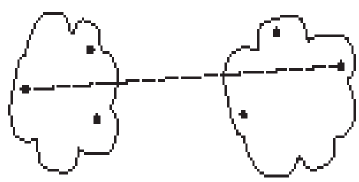
б



в



г



д

Рис.3. а – координати точок ; б – Евклидові відстані між точками; в - діаграма гніздових кластерів; г – дендрограма, кластери об'єднані за методом найбільш віддаленого сусіда (метрика Хаусдорфа); д – приклади обчислення відстаней між кластерами.

Матрична форма *р-адичного* кодування має вигляд $x = \mathbf{C}p$, де x – десяткове кодування, $C=\{c_{ij}\}$ - матриця з дендрограмними кодами, p – вектор ступенів фіксованого цілого (звичайно, більш обмеженого, фіксованого простого) p. *Р-Адичне* кодування визначене для будь-якої про′єктної множини може бути представлене як $x = \sum_{j=1}^{n-1} c_j p^j$, де $c_j \in \{-1{,}0, +1\}$ для будь-якого про′єкта *x*, пов′язаного з термінальним вузлом.

Більш детально

$$x_i = = \sum_{j=1}^{n-1} c_{ij} p^j \text{ , де } c_j \in \{-1, 0, +1\}.$$

Тут j – рівень або ранг (корінь: n-1; термінал: 1), i – про′єктний індекс. Відзначимо, що крім дендрограми (бінарного дерева) можлива також побудова орієнтованого двійкового дерева, пов′язаного з нетермінальним вузлом, яке може бути використано для характеристики НМ.

Зазначимо що p-*адичне* розширення будь-якого $q \in Q : q = \sum_{k=n}^{\infty} a_k p^k$ для деякого $n \in Z$, $a_k \in 0, 1, p-1$ для всіх $k \geq n$. Іноли q позначають цифрами, тобто $q = a_1 a_2 a_3 \; . \; . \; . \; a_r$ (так зване “backwards” – представлення).

Для ранжованої дендрограми, показаної на рис.2, застосуємо таке *p*-адичне кодування термінальних вузлів, проходячих шлях від кореня:

$$x_1 = 0 \cdot 2^7 + 0 \cdot 2^5 + 0 \cdot 2^2 + 0 \cdot 2^1;$$

$$x_2 = 0 \cdot 2^7 + 0 \cdot 2^5 + 0 \cdot 2^2 + 1 \cdot 2^1; \ldots$$

$$x_4 = 0 \cdot 2^7 + 1 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3; x_6 = \ldots$$

$$0 \cdot 2^7 + 1 \cdot 2^5 + 1 \cdot 2^4 \, ..$$

Десяткові еквіваленти p-адичного подання термінальних вузлів визначаються у вигляді: $\{x_1, x_2, ..., x_8\} =$ 0, 2, 4, 32, 40, 48, 128, 192 відповідно.

Відстані й норми визначені в наступному вигляді:

$$d_p(x, x') = d_p|x - x'| = 2^{-r+1}$$

або 2^{-r} де x=$\sum_k a_k 2^k$,

$$x' = \sum_k a'_k 2^k \,,$$

$r = \arg\min_k \{a_k = a'_k\}$. Норма визначена як $d_p(x,0) = 2^{-1+1} = = 1$.

Розмітка дендрограми, тобто маршрут від «верхівки» дендрограми аж до даних x буде «здобувати» мітки $a_\nu \in \{0{,}1\}$ на кожному рівні ν. Потім двійкове число B(x) має вигляд – $B(x) = \sum_\nu a_\nu 2^\nu$, де ν- прогони через рівні. Цей метод дозволяє використовувати 2-adic відстань на множині даних X: $d(x, y) = 2^{-\mu(x,y)}$, де $\mu(x,y) = =\max\{\nu \,|\, 2^\nu \text{ divides } B(x) - B(y)\}$ найнижчий показник, що з'являється в B(x) - B(y), розглянутий як сума степенів 2.

Несподіваний наслідок цього полягає в тому, що більші початкові значення B(x) і B(y) можуть мати в загальному випадку меншу відстань між x і y. 2-adic відстань – приклад так званої ультраметрики, і задовольняє характеристичні властивості: d(x, y) (max {d(x, z), d(z, y)} для будь-якого z є X. Відзначимо також d як p-adic метрику, де p – просте число (у нашому випадку, p = 2). Хоча загальний вибір p бере до уваги недвійковий випадок, позначаючи n гілок на даному рівні з 0, 1, . . . , p - 1, якщо p є n, проте, надають перевагу в якості p простому числу.

Великий вплив на формування портфеля має та обставина, що переважна більшість даних отримана за умов невизначеності і сформувати об’єктивну функцію належності не видається можливим. Для таких умов можливе використання бінарних ієрархічних дерев, які дозволяють врахувати функцію належності (при її використанні) або приймати рішення за умов відсутності інформації про функцію належності. Оптимізація портфеля може розглядатися в ракурсі кластеризації масивів фінансових даних і виявлення в них нових структур і пат тернів [24].

Можна зазначити , що задача вибору портфеля цінних паперів (ПЦП), яка розглядається в якості підмножини ЦП, властивості якої збігаються із властивостями вихідної множини ЦП (ринок), може бути сформульована за допомогою методів кластерного аналізу. Це позначає застосування нової стратегії формування портфеля ЦП – забезпечити поводження портфеля аналогічне (з погляду зльотів і падінь) поводженню ринку.

Висновки

1. Прийняття теорії хаосу як головної парадигми функціонування ринку припускає облік непередбачуваності ринку, отже, необхідність підходу до роботи на ринку за допомогою методів, які не ґрунтуються на прогнозах. Одна з основних концепцій сучасного розвитку ринку, як і всієї економіки, полягає в тому, що складність часто здобуває форму ієрархії, і, які ієрархічні системи мають деякі загальні властивості, що не залежать від їх специфічного змісту. Ієрархія – одна із центральних структурних схем, що складає архітектуру, яка використовується в понятті «складність».

2. Сформульовано новий клас задач – визначення портфеля ЦП, поводження якого буде схожим на поводження ринку ЦП у цілому, якщо під *поводженням* розуміти зростання і падіння вартості ЦП.

Список літератури

1. *Вильямс.б. торговый хаос. экспертные методики максимизации прибыли. интернет-ресурс http:// www.xerurg.ru*

2. Шустер Г, Детерминированный хаос, М.: Мир, 1988.

3. Федер Е., Фракталы, М.: Мир, 1991.

4. Петерс. Э. Фрактальный анализ финансовых рынков. Применение теории хаоса в инвестициях и экономике

5. Fionn Murtagh. Symmetry in Data Mining and Analysis: A Unifying View based on Hierarchy. -arXiv:0805. 2744v1 [stat.ML] 18 May 2008. – 33 pp.

6. P.E. Bradley. Mumford dendrograms. Computer Journal, 2007. submitted.

7. L. Brekke and P.G.O. Freund. p-Adic numbers in physics. Physics Reports, 233:1{66, 1993.

8. B. Dragovich and A. Dragovich. A p-adic model of DNA sequence and genetic code. Technical report, 2006. arXiv:q-bio/0607018v1.

9. F.Q. Gouvea. p-Adic Numbers: An Introduction. Springer, 2003.

10. Khrennikov. Non-Archimedean Analysis: Quantum Paradoxes, Dynamcal Systems and Biological Models. Kluwer, 1997.

11. F. Murtagh, G. Downs, and P. Contreras. Hierarchical clustering of massive, high dimensional data sets by explo-iting ultrametric embedding. SIAM Journal on Scientific Computing, 2007. In press.

12. Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений.—М.: Статистика, 1974.—240 с.

13. Дюран Н., Оделл П. Кластерный анализ.-М.: Статистика, 1977.-128 с.

14. Диде Э. и др. Методы анализа данных.-М.: Финансы и статистика, 1985.- 360 с.

15. Мандель И. Д. Кластерный анализ.- М.: Финансы и статистика. 1988.-176 с.

16. Жамбю М. Иерархический кластерный анализ и соответствия. М.: Финансы и статистика. 198.- с

17. H.A. Simon. The Sciences of the Artificial. MIT Press, Cambridge, MA, 1996.

18. В.Б . Бериков,Г.С.Лбов. Современные тенденции в кластерном анализе. – Ин-т матем. им.С.Л.Соболева СО РАН

19. Хренников А. Ю. Моделирование процессов мышления в р-адических системах координат. - М.: ФИЗМАТЛИТ, 2004.- 296 с

20. Murtag F. Thinking Ultrametrically. Інтернет-ресурс: http://

21. Murtagh F. Identifying the ultrametricity of time series. Eur. Phys. J. B ***43****, 573–579 (2005) DOI: 10.1140/epjb/e2005-00092-8*

22. Simon H.A. The Sciences of the Artificial. MIT Press, Cambridge, MA, 1996.

23. Хренников А. Ю. Моделирование процессов мышления в р-адических системах координат. - М.: ФИЗМАТЛИТ, 2004. - 296 с.

23. Мінаєва Ю.І. Методи нечіткої ієрархічної кластеризації в задачах формування інвестиційного портфелю акцій: Матеріали міжнародної научної конференції «Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта (isdmci'2010)»

24. Мінаєва Ю.І. Визначення портфеля цінних паперів на основі інваріантних підмножин ієрархічної кластеризації. Системний аналіз та інформаційні технології: Матеріали XII Міжнародної науково-технічної конференції (25-29 травня 2010 р., Київ). – К.: ННК «ІПСА» НТУУ «КПІ», 2010. – 290 с.

25. Мінаєва Ю. І. Кластерний аналіз в задачах вибору портфелю цінних паперів. Наукова конференція молодих вчених, аспірантів і студентів КНУБА (16-18 листопада 2009р., Київ) : тези доповідей. – в 2-х частинах. – Ч.1. - К.: КНУБА, 2010. – 157 с.

Стаття надійшла до редколегії:12.11.2010

Рецензент: д-р техн. наук, проф. М.К. Печурін, Національний авіаційний університет, Київ.