# МЕТОДИКА АДАПТИВНОЙ КЛАСТЕРИЗАЦИИ

# ФАКТОГРАФИЧЕСКИХ ДАННЫХ

## Цель работы:

Повышение эффективности процесса кластеризации фактографических данных с использованием методов интеллектуального анализа данных.

## Назначение работы:

Оптимизация списка услуг брокерского обслуживания для выделенных групп клиентов.

## Задачи, подлежащие решению:

1. Исследование методов интеллектуального анализа данных, используемых для кластеризации фактографических данных.

2. Исследование деятельности кредитной организации с целью выявления разновидностей показателей, характеризующих бизнес – процессы.

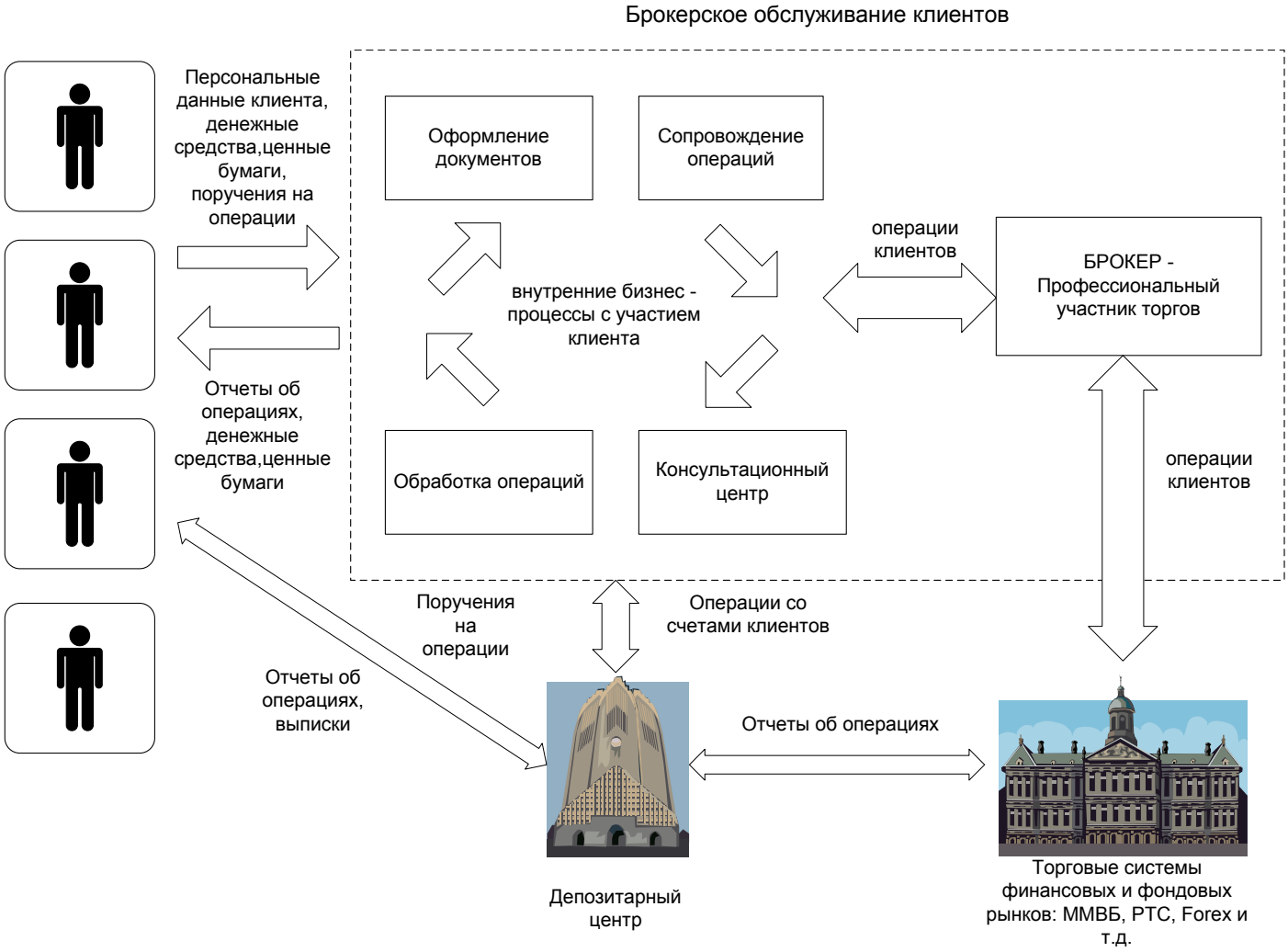
3. Разработка критериев оценки/выбора исследованных методов кластерного анализа. Обобщение исходных данных алгоритмов с целью оценки возможности адаптации алгоритмов к различным предметным областям.

4. Разработка адаптивного алгоритма кластеризации фактографических данных.

5. Разработка программного комплекса интеллектуального анализа данных.

6. Апробация методов и анализ экспериментальных данных.

# ОПИСАНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ




# СУЩЕСТВУЮЩИЕ РЕШЕНИЯ И ИХ НЕДОСТАТКИ:

| **Продукт/ Характеристики** | **Информационно – аналитическая система «Арион»** | **Информационно – аналитическая система «Прогноз»** | **Аналитическая платформа «Deductor»** | **Business Object Enterprise** | **R&K Cognos Analyst** |
|---|---|---|---|---|---|
| **Тип системы** | Информационно - аналитическая система | Информационно - аналитическая система | Инструментарий для создания на базе аналитической платформы эффективной аналитической системы | Инструментарий для формирования аналитической отчетности | Инструментарий для построения системы принятия решения |
| **Функциональное назначение** | Обработка и представление информации на естественном языке, выявление логических взаимосвязей и развернутый поиск | Аналитическая система поддержки принятия решения с использованием имитационных, оптимизационных и статистических методов | Корпоративная отчетность, прогнозирование, сегментация, поиск закономерностей | Корпоративная отчетность | Анализ финансового состояния предприятия |
| **Недостатки** | отсутствие комплексного подхода при обработке данных<br>узкая специализация на предметной области небольшое количество реализованных методов интеллектуального анализа данных | | | узкая специализация на предметной области небольшое количество реализованных методов интеллектуального анализа данных | |

# ФОРМАЛИЗОВАННАЯ МОДЕЛЬ ПРЕДМЕТНОЙ ОБЛАСТИ

$\mathbf{Cl} = \{ Cl^{Fiz}, Cl^{Jur} \}$, $|Cl^{Fiz}| \geq 1$, $|Cl^{Jur}| \geq 1$, $\mathbf{Cl^{Fiz}} = \{ Cl^{Fiz}_{PersInfo}, Cl^{Fiz}_{Type}, Cl^{Fiz}_{ListUslug} \}$, $\mathbf{Cl^{Jur}} = \{ Cl^{Jur}_{PersInfo}, Cl^{Jur}_{Type}, Cl^{Jur}_{ListUslug} \}$

$\mathbf{Cl^{Fiz}_{PersInfo}} = \{$ INN,FIO,FIO1,FIO2,Brief,Country,BirthDate,RegDocument,FlResident,Sex,FizAddress,PostAddress,Telefon,EMail$\}$

$\mathbf{Cl^{Jur}_{PersInfo}} = \left\{ \begin{array}{l} \text{INN,Name,Name1,Name2,Brief,Country,JurType,BankRecviz,JurState,RegDocument,FlResident,FizAddress,PostAddress,} \\ \text{Telefon,EMail} \end{array} \right\}$

RegDocument $= \{$ DocType,Series,Number,Date,RegPlace$\}$, DocType $= \{$ Паспорт РФ,Свидетельство о регистрации$\}$, FlResident $= \{$ Да, Нет$\}$

JurType $= \{$ ЮР.лицо, Банк, РКЦ, Отделение банка$\}$, Sex $= \{$ Женский, Мужской$\}$, BankRecviz $= \{$ BIK, CorAccount, Account, BankName$\}$

JurState $= \{$ Головная организация, Филиал, Дочерняя организация, Зависимая компания, Подразделение $\}$

$\mathbf{Uslugi} = \{ Op^{UsualDealB}, Op^{UsualDealV}, Op^{RepoDealB}, Op^{RepoDealV}, Op^{InSecurities}, Op^{OutSecurities}, Op^{MoveSecurities}, Op^{MargDeal}, Op^{InMoney}, Op^{OutMoney}, Op^{MoveMoney} \}$

$\mathbf{Op^{UsualDealB}} = \{$ DealType, DealDate, TradeTime, TradeNo, Security, DealNum, DealPrice, DealQty, DealCoupon, OrderNo, TradePlace, Owner, Portfolio$\}$

$\mathbf{Op^{UsualDealV}} = \left\{ \begin{array}{l} \text{DealType, DealDate, TradeTime, DogovNo, Security, DealNum, DealPrice, DealQty, DealCoupon,} \\ \text{DealFund, PayFund, DelDate, PayDate, Owner, Portfolio} \end{array} \right\}$

$\mathbf{Op^{RepoDealB}} = \left\{ \begin{array}{l} \text{DealType, DealDate, TradeTime, TradeNo, Security, DealNum, DealPrice1, DealQty1, DealCoupon1,} \\ \text{DealPrice2, DealQty2, DealCoupon2, OrderNo, TradePlace, RepoPers, DelDate1, PayDate1, DelDate2, PayDate2, Owner, Portfolio} \end{array} \right\}$

$\mathbf{Op^{RepoDealV}} = \left\{ \begin{array}{l} \text{DealType, DealDate, TradeTime, DogovNo, Security, DealNum, DealPrice1, DealQty1, DealCoupon1,} \\ \text{DealPrice2, DealQty2, DealCoupon2, RepoPers, DelDate1, PayDate1, DelDate2, PayDate2, DealFund, PayFund, Owner, Portfolio} \end{array} \right\}$

$\mathbf{Op^{InSecurities}} = \{$ OperDate, OperTime, OperNo, Security, Num, Owner, Portfolio, DelDate, Depo1, DepoAcc1, DepoRazd1$\}$

$\mathbf{Op^{OutSecurities}} = \left\{ \begin{array}{l} \text{OperDate, OperTime, OperNo, Security, Num, Owner, Portfolio, DelDate, Depo1, DepoAcc1, DepoRazd1,} \\ \text{Depo2, DepoAcc2, DepoRazd2} \end{array} \right\}$

$\mathbf{Op^{MoveSecurities}} = \left\{ \begin{array}{l} \text{OperDate, OperTime, OperNo, Direction, Security, Num, Owner, Portfolio, DelDate,} \\ \text{Depo1, DepoAcc1, DepoRazd1, Depo2, DepoAcc2, DepoRazd2} \end{array} \right\}$

$\mathbf{Op^{MargDeal}} = \left\{ \begin{array}{l} \text{DealType, DealDate, TradeTime, TradeNo, Security, DealNum, DealPrice, DealQty, DealCoupon,} \\ \text{OrderNo, TradePlace, Owner, Portfolio, MargValue1, MargValue2} \end{array} \right\}$

$\mathbf{Op^{InMoney}} = \{$ OperDate, OperTime, OperNo, Fund, Qty, Owner, Portfolio, PayDate, BankRecviz1$\}$ , где BankRecviz1 $\in \{$ BankRecviz$\}$

$\mathbf{Op^{OutMoney}} = \{$ OperDate, OperTime, OperNo, Fund, Qty, Owner, Portfolio, PayDate, BankRecviz1,BankRecviz2$\}$ , где

$\{$ BankRecviz1, BankRecviz2$\} \subset \{$ BankRecviz$\}$

$\mathbf{Op^{MoveMoney}} = \{$ OperDate, OperTime, OperNo, Direction, Fund, Qty, Owner, Portfolio, PayDate, BankRecviz1,BankRecviz2$\}$ , где

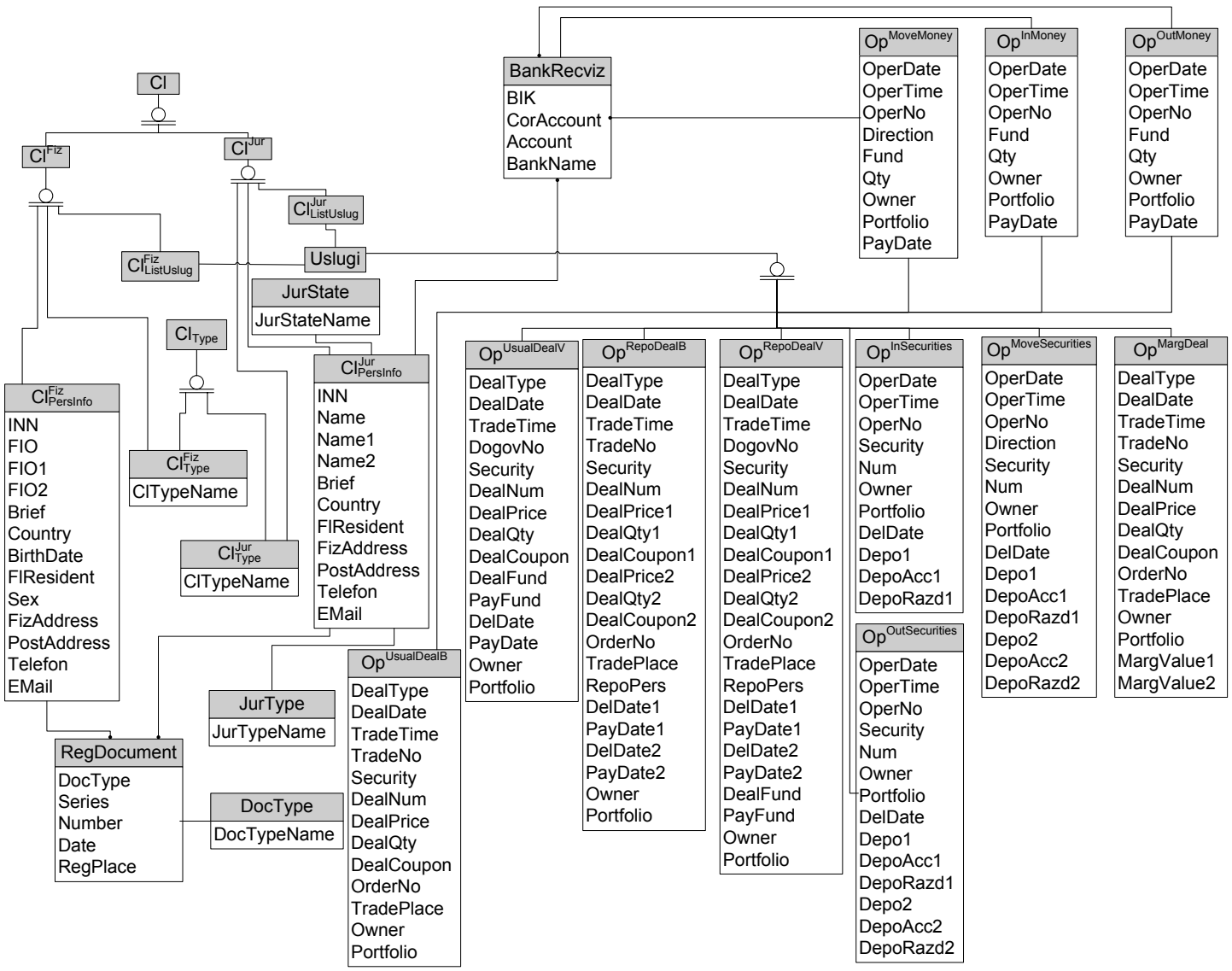
$\{$ BankRecviz1, BankRecviz2$\} \subset \{$ BankRecviz$\}$

$\mathbf{Cl^{Fiz}_{ListUslug}} = \{ U_i \}_{n_1}$, $U_i \in$ Uslugi, $n_1 \geq 1$, $\mathbf{Cl^{Jur}_{ListUslug}} = \{ U_i \}_{n_2}$, $U_i \in$ Uslugi, $n_2 \geq 1$, $\mathbf{Cl_{ListUslug}} = \{ Cl^{Fiz}_{ListUslug}, Cl^{Jur}_{ListUslug} \}$, $|Cl_{ListUslug}| \geq 1$

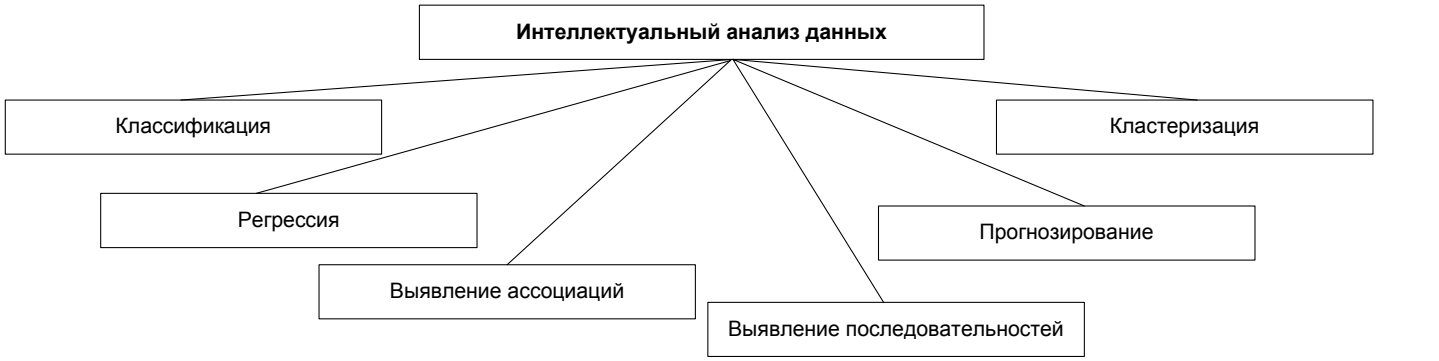
$\mathbf{Cl^{Fiz}_{Type}} = \{ Cl^{Fiz}_{Type_1}, Cl^{Fiz}_{Type_2}, \ldots, Cl^{Fiz}_{Type_m} \}$ , $m \geq 0$, $\mathbf{Cl^{Jur}_{Type}} = \{ Cl^{Jur}_{Type_1}, Cl^{Jur}_{Type_2}, \ldots, Cl^{Jur}_{Type_k} \}$ , $k \geq 0$, $\mathbf{Cl_{Type}} = Cl^{Fiz}_{Type} \cup Cl^{Jur}_{Type}$ , $|Cl_{Type}| \geq 1$

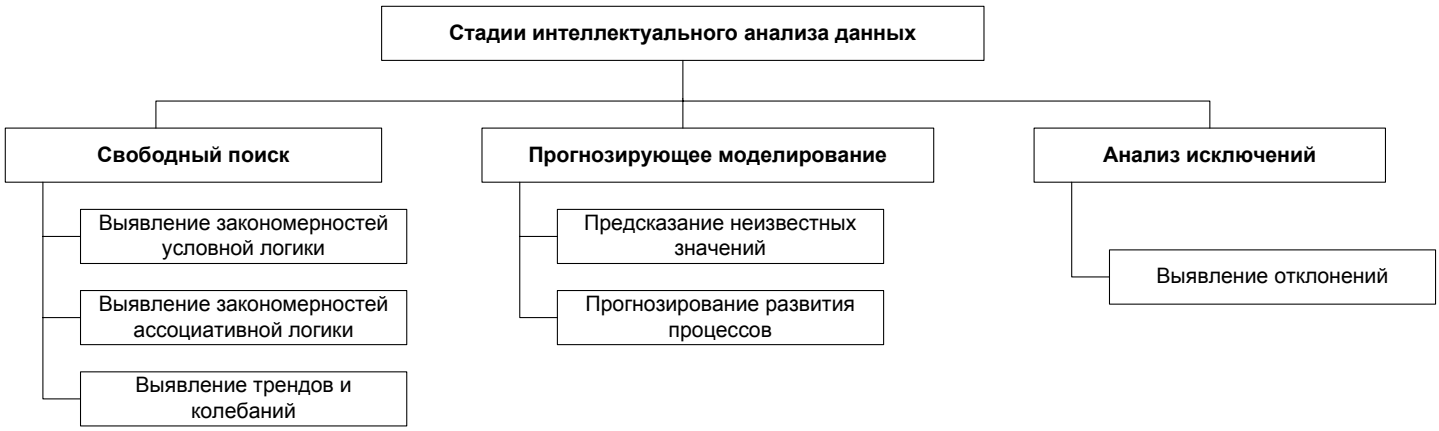
$\mathbf{Cl_{Type}} = \gamma \left( \{ \langle Cl_{PersInfo}, Cl_{ListUslug} \rangle \} \right)$, $\boldsymbol{\gamma} = \{ \gamma_1, \gamma_2, \ldots, \gamma_1 \}$ , $1 > 0$

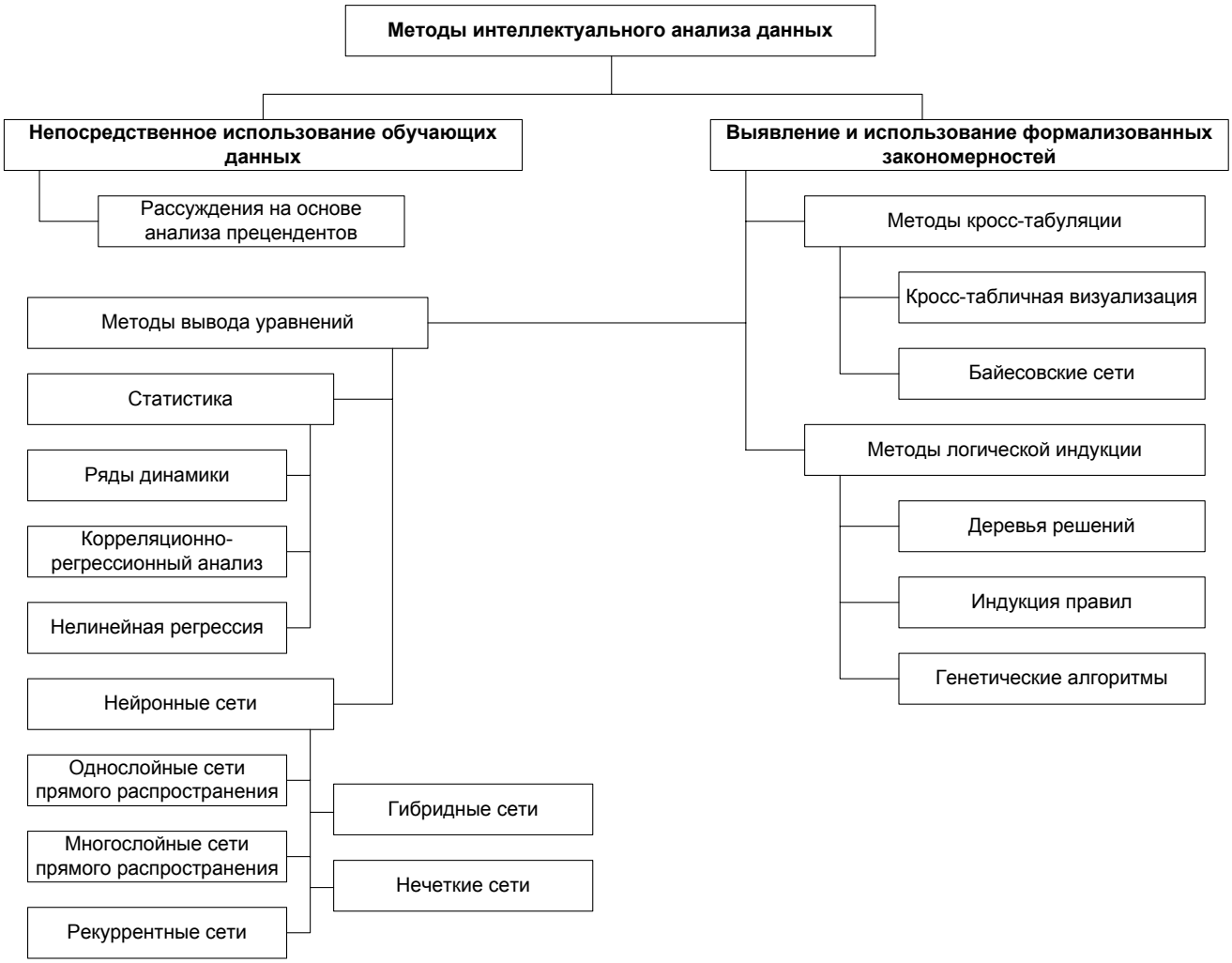
$\mathbf{n_1}$ - ? $\mathbf{n_2}$ - ? $\mathbf{n}$ - ? $\mathbf{m}$ - ? $\mathbf{k}$ - ? $\mathbf{Cl^{Fiz}_{Type}}$ - ? $\mathbf{Cl^{Jur}_{Type}}$ - ? $\mathbf{Cl_{Type}}$ - ? $\boldsymbol{\gamma_1 \bullet \gamma_2 \bullet \ldots \bullet \gamma_1}$ - ?, $\boldsymbol{\varphi}\left( \mathbf{Cl_{ListUslug}} \middle| \mathbf{Cl_{Type}}, \mathbf{Доходность, Экон. риски} \right) = \text{optimum}, \gamma \rightarrow \varphi$



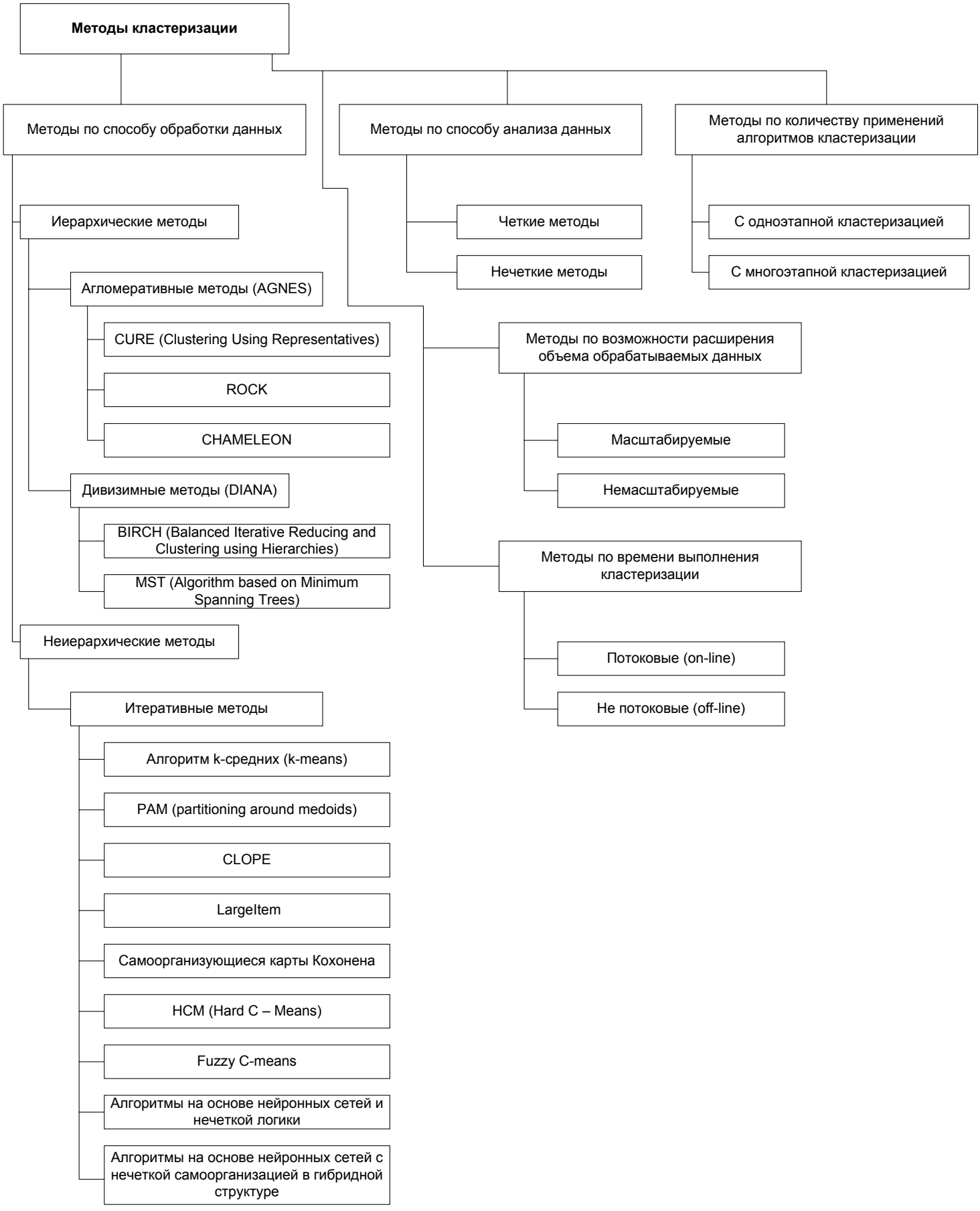
# КЛАССИФИКАЦИЯ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ



# КЛАССИФИКАЦИЯ МЕТОДОВ КЛАСТЕРИЗАЦИИ



# КРИТЕРИИ ВЫБОРА МЕТОДОВ КЛАСТЕРИЗАЦИИ

Критерий 1 ($Cr_1$): Объем информации (количество строчек или по занимаемому месту) по отношению к времени обработки.

Критерий 2 ($Cr_2$): Размерность информации (количество атрибутов в строке) в порядковом выражении.

Критерий 3 ($Cr_3$): Типы атрибутов: числовой дискретный, числовой непрерывный, строковый.

Критерий 4 ($Cr_4$): Чувствительность к равномерности информации (наличие аномалий-выбросов во входном наборе данных).

Критерий 5 ($Cr_5$): Априорное (экспертное) представление о форме получаемых кластеров.

Критерий 6 ($Cr_6$): Априорное (экспертное) представление о количестве кластеров.

Критерий 7 ($Cr_7$): Априорное (экспертное) представление о перекрываемости кластеров.

Критерий 8 ($Cr_8$): Качество кластеризации.

| Критерий / Метод | **Критерий 1** ($Cr_1$) | **Критерий 2** ($Cr_2$) | **Критерий 3** ($Cr_3$) | **Критерий 4** ($Cr_4$) | **Критерий 5** ($Cr_5$) | **Критерий 6** ($Cr_6$) | **Критерий 7** ($Cr_7$) | **Критерий 8** ($Cr_8$) |
|---|---|---|---|---|---|---|---|---|
| CURE | линейная зависимость, большой объем данных | невысокая размерность | числовые | нечувствителен | сложная форма | входной параметр | без пересечений | высокое |
| BIRCH | линейная зависимость, большой объем данных | невысокая размерность | числовые | чувствителен | сферическая форма | входной параметр | без пересечений | среднее |
| MST | линейно-логарифмическая зависимость, большой объем данных | средняя размерность | числовые | чувствителен | сложная форма, в т.ч. выпуклой и вогнутой форм | входной параметр | без пересечений | высокое |
| k-средних | нелинейная зависимость, небольшой объем данных | низкая размерность | числовые | очень чувствителен | сферическая форма | входной параметр | без пересечений | низкое |
| PAM | нелинейная зависимость, небольшой объем данных | низкая размерность | числовые | чувствителен | сложная форма | входной параметр | без пересечений | среднее |
| CLOPE | линейная зависимость, огромный объем данных | средняя размерность | числовые | чувствителен | сложная форма | вычисляемая величина | без пересечений | высокое |
| Самоорганизующиеся карты Кохонена | нелинейная зависимость, большой объем данных | высокая размерность | числовые | низкая чувствительность | сложная форма | входной параметр | без пересечений | среднее |
| HCM | нелинейная зависимость, большой объем данных | средняя размерность | числовые | чувствителен | сложная форма | входной параметр | без пересечений | среднее |
| Fuzzy C-means | нелинейная зависимость, большой объем данных | высокая размерность | числовые | чувствителен | сложная форма | входной параметр | с пересечениями | среднее |

# АДАПТАЦИЯ МЕТОДОВ КЛАСТЕРИЗАЦИИ

$Par^{Birch}_{Dist}$ - способ определения дистанции между кластерами

$Par^{Birch}_{Qual}$ - метод оценки качества кластеризации

$Par^{Birch}_{ThrQual}$ - порог для метода оценки качества кластеризации

$Par^{Birch}_{Thr1}$ - начальный порог (Фаза 1)

$Par^{Birch}_{OutlinePerc1}$ - процент аномалий (выбросов) в общем объеме (Фаза 1)

$Par^{Birch}_{ClustAlg3}$ - алгоритм выполнения дополнительной кластеризации (Фаза 3)

$Par^{Birch}_{RefPass4}$ - количество очисток кластеров (Фаза 4)

$Par^{Birch}_{OutlinePerc4}$ - процент аномалий (выбросов) в общем объеме (Фаза 4)

$Par^{Birch}_{k}$ - количество кластеров

$Par^{HCM}_{c}$ - количество кластеров

$Par^{HCM}_{StopThr}$ - пороговая величина для остановки алгоритма

$Par^{PAM}_{k}$ - количество кластеров

$Par^{PAM}_{MethodCh}$ - способ выбора начальных центров

$Par^{PAM}_{MaxIterNum}$ - максимальное количество итераций

$Par^{Fuzzy\ C\text{-}means}_{c}$ - количество кластеров

$Par^{Fuzzy\ C\text{-}means}_{m}$ - экспонентциальный вес

$Par^{Fuzzy\ C\text{-}means}_{StopThr}$ - пороговая величина для остановки алгоритма

$Par^{k\text{-средних}}_{k}$ - количество кластеров

$Par^{k\text{-средних}}_{MethodCh}$ - способ выбора начальных центров

$Par^{k\text{-средних}}_{MaxIterNum}$ - максимальное количество итераций

$Par^{MST}_{k}$ - количество кластеров

$Par^{Kohonen}_{k}$ - количество кластеров

$Par^{Kohonen}_{VolPerc1}$ - объем обучающего множества

$Par^{Kohonen}_{VolPerc2}$ - объем валидационного множества

$Par^{Kohonen}_{VolPerc3}$ - объем тестового множества

$Par^{Kohonen}_{DivMethod}$ - способ разделения множества

$Par^{Kohonen}_{StopThr}$ - порог остановки алгоритма обучения и тестирования нейронов

$Par^{Kohonen}_{TrainRate}$ - скорость обучения нейронов

$Par^{Clope}_{k}$ - количество кластеров

$Par^{Cure}_{k}$ - количество кластеров

$Par^{Cure}_{c}$ - количество точек, которое захватывается при обработке за один раз

$Par^{Cure}_{p}$ - количество предварительных разделов

$Par^{Cure}_{q}$ - сокращающий коэффициент для предварительных разделов

$Par^{Cure}_{a}$ - коэффициент удаленности

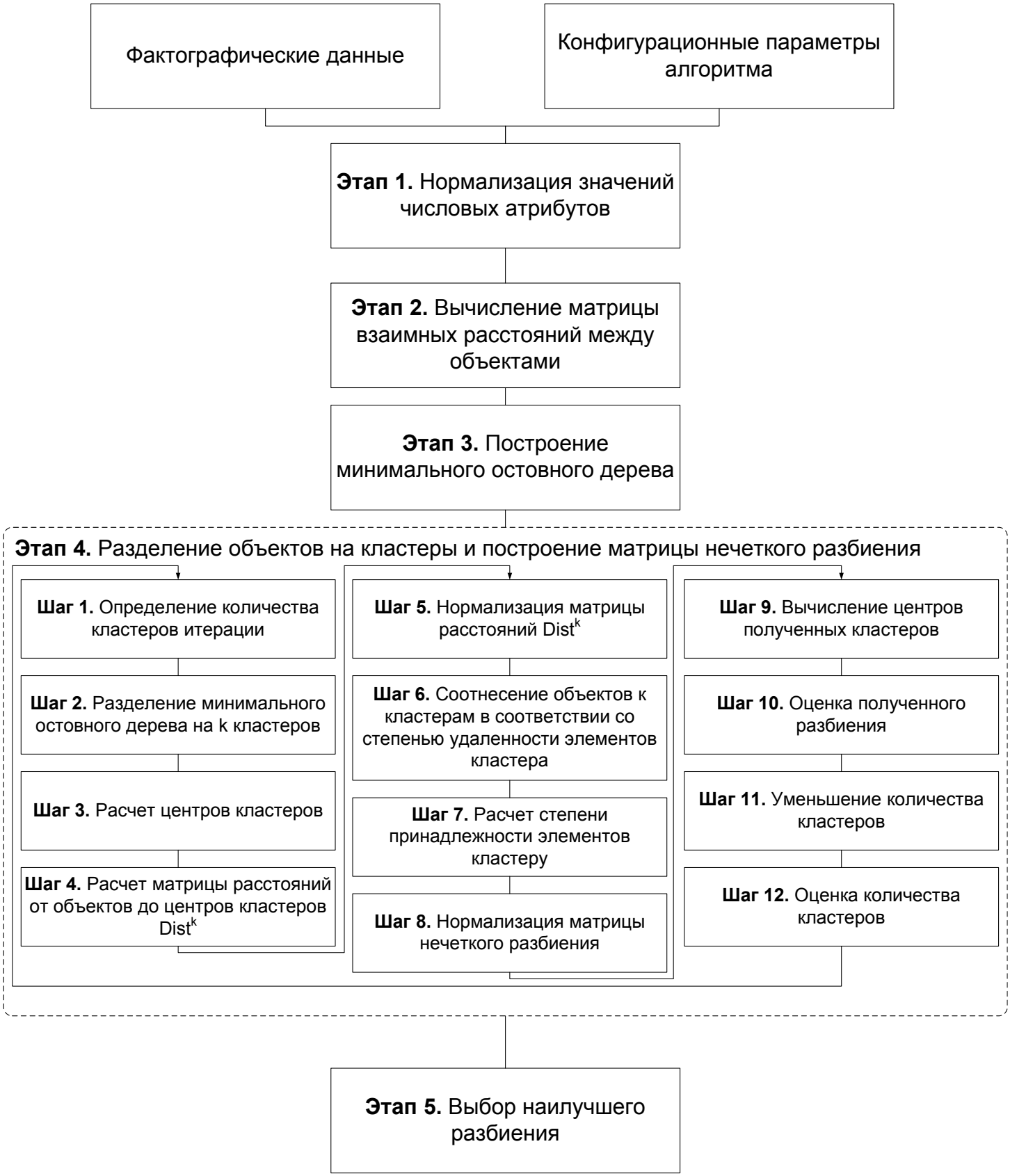
$$Par^{Cure}_{k} \equiv Par^{Birch}_{k} \equiv Par^{MST}_{k} \equiv Par^{k\text{-средних}}_{k} \equiv Par^{PAM}_{k} \equiv Par^{Kohonen}_{k} \equiv Par^{HCM}_{c} \equiv Par^{Fuzzy\ C\text{-}means}_{c} \equiv Par_{c}$$

$$Par^{Kohonen}_{VolPerc1} + Par^{Kohonen}_{VolPerc2} + Par^{Kohonen}_{VolPerc3} = 100\% \; ; \; Par^{k\text{-средних}}_{MethodCh} \equiv Par^{PAM}_{MethodCh} \equiv Par_{MethodCh} \; ; \; Par^{k\text{-средних}}_{MaxIterNum} \equiv Par^{PAM}_{MaxIterNum} \equiv Par_{MaxIterNum}$$

| Параметр / Метод | Параметр 1 | Параметр 2 | Параметр 3 | Параметр 4 | Параметр 5 | Параметр 6 | Параметр 7 | Параметр 8 | Параметр 9 |
|---|---|---|---|---|---|---|---|---|---|
| CURE | $Par^{Cure}_{k}$ | $Par^{Cure}_{c}$ | $Par^{Cure}_{q}$ | $Par^{Cure}_{a}$ | $Par^{Cure}_{p}$ | | | | |
| BIRCH | $Par^{Birch}_{k}$ | $Par^{Birch}_{Dist}$ | $Par^{Birch}_{Qual}$ | $Par^{Birch}_{ThrQual}$ | $Par^{Birch}_{Thr1}$ | $Par^{Birch}_{OutlinePerc1}$ | $Par^{Birch}_{ClustAlg3}$ | $Par^{Birch}_{RefPass4}$ | $Par^{Birch}_{OutlinePerc4}$ |
| MST | $Par^{MST}_{k}$ | | | | | | | | |
| k-средних | $Par^{k\text{-средних}}_{k}$ | $Par^{k\text{-средних}}_{MethodCh}$ | $Par^{k\text{-средних}}_{MaxIterNum}$ | | | | | | |
| PAM | $Par^{PAM}_{k}$ | $Par^{PAM}_{MethodCh}$ | $Par^{PAM}_{MaxIterNum}$ | | | | | | |
| CLOPE | $Par^{Clope}_{r}$ | | | | | | | | |
| Само-организующиеся карты Кохонена | $Par^{Kohonen}_{k}$ | $Par^{Kohonen}_{VolPerc1}$ | $Par^{Kohonen}_{VolPerc2}$ | $Par^{Kohonen}_{VolPerc3}$ | $Par^{Kohonen}_{DivMethod}$ | $Par^{Kohonen}_{StopThr}$ | $Par^{Kohonen}_{TrainRate}$ | | |
| HCM | $Par^{HCM}_{c}$ | $Par^{HCM}_{StopThr}$ | | | | | | | |
| Fuzzy C-means | $Par^{Fuzzy\ C\text{-}means}_{c}$ | $Par^{Fuzzy\ C\text{-}means}_{m}$ | $Par^{Fuzzy\ C\text{-}means}_{StopThr}$ | | | | | | |

# АДАПТИВНЫЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ ФАКТОГРАФИЧЕСКИХ ДАННЫХ

## Укрупненная схема



# АДАПТИВНЫЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ ФАКТОГРАФИЧЕСКИХ ДАННЫХ (Этапы 1 - 4)

*Входные данные алгоритма:*

$D = \{u_1, u_2, ..., u_m\}$, где $u_i$ – объекты кластеризации, $m$ – количество объектов кластеризации, $i = \overline{1,m}$;

$u_i = \{(Value_{i1}, t_1), (Value_{i2}, t_2), ..., (Value_{in}, t_n)\}$, где $Value_{ij}$ – значение $j^{го}$ атрибута $i^{го}$ объекта кластеризации, $t_j$ – тип атрибута объекта кластеризации, $n$ – количество атрибутов объекта кластеризации, $j = \overline{1,n}$;

$t_j = \{ValueType_j, FieldType_j\}$, где $ValueType_j$ – тип значения атрибута, $ValueType_j \in ValueTypes$, $FieldType_j$ – вид значения атрибута, $FieldType_j \in FieldTypes$; $FieldTypes = \{Входное, Идентифицирующее, Информационное\}$;

$ValueTypes = \{Целочисленный\ тип, Денежный\ тип, Лингвистический\ тип\}$, где $Целочисленный\ тип \subset \mathbb{Z}$,

$Денежный\ тип \subset R$, $Лингвистический\ тип \subset Словарная\ система$;

$Словарная\ система = \{Лингв.тип1, Лингв.тип2, ..., Лингв.типS\}$, где $Лингв.тип_i$ – объект словарной системы, характеризующий оценочные/качественные показатели объектов кластеризации;

$q$ – максимальное количество кластеров, $q \le m$; $p$ – размазанность кластеров, $p \in (0;10]$;

$K = \{K_1, K_2, ..., K_n\}$, где $K_i$ – весовой коэффициент влияния атрибута объекта, $K_i \in [0;1]$;

$w$ – степень удаленности элементов, $w \in (0;1]$;

$Metric$ – способ определения расстояния между объектами, $Metric \in Metrics$;

$$Metrics = \begin{Bmatrix} Евклидово\ расстояние, Квадрат\ Евклидова \\ расстояния, расстояние Чебышева \end{Bmatrix};$$

$OstTreeMethod$ – способ построения минимального остовного дерева, $OstTreeMethod \in OstTreeMethods$;

$$OstTreeMethods = \begin{Bmatrix} Алгоритм\ Борувки, Алгоритм \\ Крускала,\ \ Алгоритм\ Прима \end{Bmatrix};$$

$NormMethod$ – способ проведения нормализации значений числовых атрибутов, $NormMethod \in NormMethods$;

$NormMethods = \{Линейная\ нормализация, Статистическая\ нормализация\}$

*Выходные данные алгоритма:* $C = \left\{C_1, C_2, ..., C_c \middle| O^c \to max, c \le q, C_1 \cup C_2 \cup ... \cup C_c = D\right\}$, $u_i \overset{\mu_{ij} \in [0,1]}{\in} C_j$, $i = \overline{1,m}$, $j = \overline{1,c}$

*Описание алгоритма:*

**Этап 1.** Нормализация значений числовых атрибутов.

В случае линейной нормализации выполняется следующее:

$$Value_{ij} := \left\{ \frac{Value_{ij}}{\underset{j}{Max}(Value_{ij})} \middle| \begin{matrix} \underset{j}{Max}(Value_{ij}) \ne 0, \\ t_j \in \{Целочисленный\ тип, Денежный\ тип\} \end{matrix} \right\}$$

В случае статистической нормализации выполняется следующее:

$$Value_{ij} := \left\{ \frac{Value_{ij} - \frac{\sum_{i=1}^{m} Value_{ij}}{m}}{\sqrt{\frac{\sum_{i=1}^{m} (Value_{ij})^2}{m} - \left(\frac{\sum_{i=1}^{m} Value_{ij}}{m}\right)^2}} \middle| \begin{matrix} \underset{j}{Max}(Value_{ij}) \ne 0, \\ t_j \in \begin{Bmatrix} Целочисленный\ тип, \\ Денежный\ тип \end{Bmatrix} \end{matrix} \right\}$$

**Этап 2.** Вычисление матрицы взаимных расстояний между объектами.

$Dist_{ij} = \|u_i - u_j\| = Metric(u_i, u_j)$, где $Metric$ – способ определения расстояния между объектами.

Если $Metric = Евклидово\ расстояние$, то $Dist_{ij} = \sqrt{\sum_w \left(\left[Value_{iw} - Value_{jw}\right] * K_w\right)^2}$,

Если $Metric = Квадрат\ Евклидова\ расстояния$, то $Dist_{ij} = \sum_w \left(\left[Value_{iw} - Value_{jw}\right] * K_w\right)^2$,

Если $Metric = Расстояние Чебышева$, то $Dist_{ij} = \underset{w}{Max}\left[Value_{iw} - Value_{jw}\right] * K_w$,

где $i, j \in [1,m], w = \overline{1,n}$ при условии $FieldType[w] = "Входное"$

**Этап 3.** Построение минимального остовного дерева по выбранному способу построения дерева с использованием матрицы взаимных расстояний между объектами $Dist$.

**Этап 4.** Разделение объектов на кластеры и построение матрицы нечеткого разбиения F.

Матрица нечеткого разбиения: $F = [\mu_{ij}], \mu_{ij} \in [0,1], i \le q, j = \overline{1,m}$, где $\mu_{ij}$ – степень принадлежности $i^{го}$ объекта к $j^{му}$ кластеру. Матрица разбиения обладает следующими свойствами: $\sum_{i=1}^{k} \mu_{ij} = 1, j = \overline{1,m}$, $0 < \sum_{j=1}^{m} \mu_{ij} \le m, i = \overline{1,k}$.

# АДАПТИВНЫЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ ФАКТОГРАФИЧЕСКИХ ДАННЫХ (Этапы 4 - 5)

**Шаг 1.** Определение количества кластеров итерации: $k := q$.

**Шаг 2.** $Dist_{ij}^{k} := \left\{ 0 \left| Dist_{ij}^{k} \overset{i,j=\overline{1,m}}{=} Max \right. \right\}$.

**Шаг 3.** Расчет центров выделенных кластеров $V_i^k$.

$V_i^k = Avg\left(\left\{u_j \left| u_j \in C_i^k \right.\right\}\right)$, где $Avg$ – оператор вычисления среднего значения показателей объектов, входящих в кластер $k$, $i = \overline{1,k}$, $j = \overline{1,m}$.

Для числовых типов оператор $Avg$ определяется выражением:

$$Avg[r] = \frac{\sum\limits_{u_j \in V_i^k} \left\{ Value_{jr} \left| FieldType[w] = "Входное" \right. \right\}}{\left| V_i^k \right|},\ j = \overline{1,m},\ r = \overline{1,n}.$$

Для лингвистических типов оператор $Avg$ определяется выражением:

$$Avg[r] = \left\{ \underset{\sum \left\| Value_{jr} - Value_{lr} \right\| = min \text{ или } \underset{u_j \in V_i^k, u_l \in V_i^k}{\varphi = max}}{Value_{jr}} \left| FieldType[w] = "Входное" \right. \right\},\ r = \overline{1,n},\ j = \overline{1,m},\ l = \overline{1,m},$$ где $\varphi$ – частота значения атрибута в пределах кластера $V_i^k$.

**Шаг 4.** Расчет матрицы расстояний от объектов до центров кластеров $V_i^k$: $Dist_{ij}^{k} = \left\| V_i^k - u_j \right\| = Metric\left(V_i^k, u_j\right)$, $i = \overline{1,k}$, $j = \overline{1,m}$, где $Metric$ – способ определения расстояния между объектами.

**Шаг 5.** Нормализация матрицы расстояний от объектов до центров кластеров $V_i^k$:

$$Dist_{ij}^{k'} = \begin{cases} \dfrac{Dist_{ij}^{k}}{Max\left(Dist_{ij}^{k}\right)}, Max\left(Dist_{ij}^{k}\right) \neq 0 \\ 1, Max\left(Dist_{ij}^{k}\right) = 0 \end{cases},\ i = \overline{1,k},\ j = \overline{1,m}.$$

**Шаг 6.** Соотнесение объектов к кластерам в соответствии со степенью удаленности элементов кластера ( $w$ ):

$u_j \in V_i^k \left| Dist_{ij}^{k'} \leq w \text{ или } Dist_{ij}^{k'} = \underset{i}{Min}\left(Dist_{ij}^{k'}\right),\ i = \overline{1,k},\ j = \overline{1,m}.\right.$

**Шаг 7.** Расчет степени принадлежности кластеру: $\mu_{ij} = \left(1 - Dist_{ij}^{k'}\right)^2$, $i = \overline{1,k}$, $j = \overline{1,m}$.

**Шаг 8.** Нормализация матрицы нечеткого разбиения: $\mu_{ij} = \dfrac{\mu_{ij}}{\sum\limits_{i=1}^{k} \mu_{ij}}$, $j = \overline{1,m}$.

**Шаг 9.** Вычисление центров полученных кластеров с использованием матрицы нечеткого разбиения:

$V_i^{k'} = \dfrac{\sum\limits_{j=1}^{m} {\mu_{ij}}^p * u_j}{\sum\limits_{\substack{j=1 \\ u_j \in V_i^{k'}}}^{m} {\mu_{ij}}^p}, i = \overline{1,k}$. Для лингвистических атрибутов центра кластера вычисление производится с использованием выражения: $V_i^{k'}[r] = \underset{\mu_{ij} = Max\left(\mu_{ij}\right)}{Value_{jr}}$.

**Шаг 10.** Оценка качества разбиения: $O^k = \dfrac{\sum\limits_{i=\overline{1,k}} \dfrac{\left|V_i^{k'}\right| * \sum\limits_{j=1}^{m} {\mu^p}_{ij} * \underset{u_j \in V_i^{k'}}{\left\| V_i^{k'} - u_j \right\|}}{\underset{i \neq j}{\underset{u_j \in V_i^{k'}}{Min}\left(\left\| V_i^{k'} - u_j \right\|\right)} * \underset{u_j \in V_i^{k'}}{Max\left(\left\| V_i^{k'} - u_j \right\|\right)} * \sum\limits_{j=1}^{m} \underset{u_j \in V_i^{k'}}{\left\| V_i^{k'} - u_j \right\|} * k}}{m * k^2}$, где

$\left|V_i^{k'}\right|$ – количество элементов в кластере $i$;

$\left\| V_i^{k'} - u_j \right\| = Metric\left(V_i^{k'}, u_j\right)$ – расстояние от центра кластера $i$ до элемента $u_j$;

$u_j \in V_i^{k'}$ – отражение условия о принадлежности элемента кластеру.

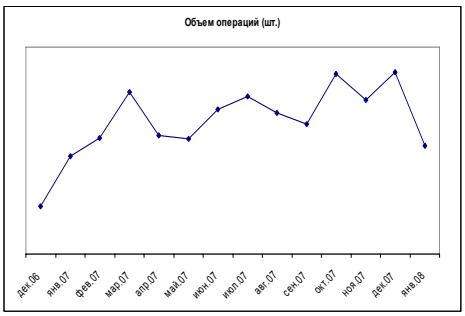
**Шаг 11.** $k := k - 1$.

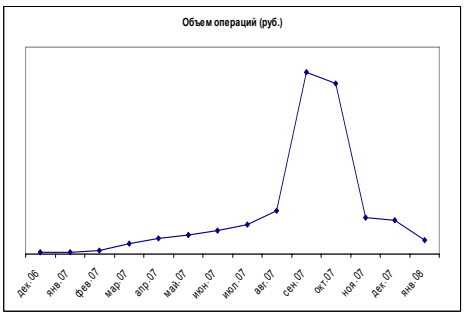
**Шаг 12.** Если $k > 0$, то переход на шаг 2.

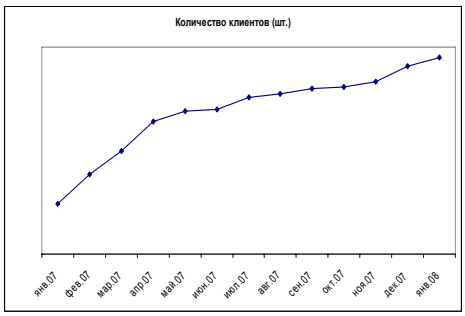
**Этап 5.** Выбор наилучшего разбиения: $O_{Onm} = \underset{i=\overline{1,q}}{MAX}\left(O^i\right)$.

# АНАЛИЗ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

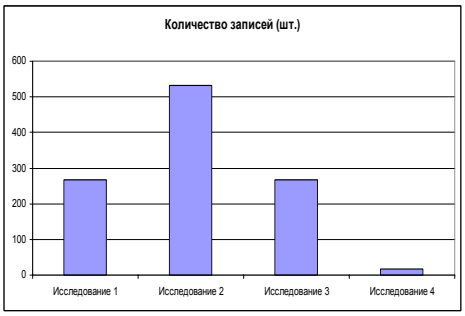
## Показатели операций брокерского обслуживания

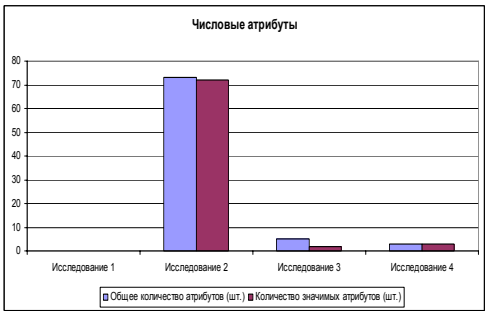


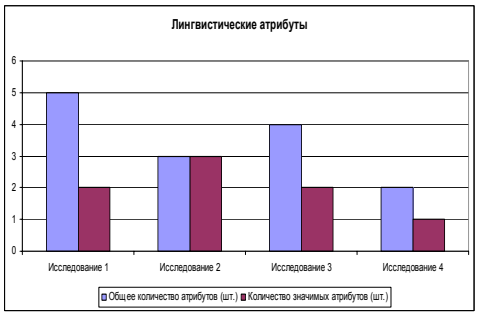



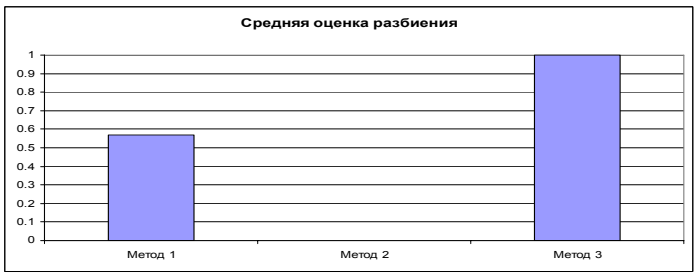

## Характеристики исследуемых данных



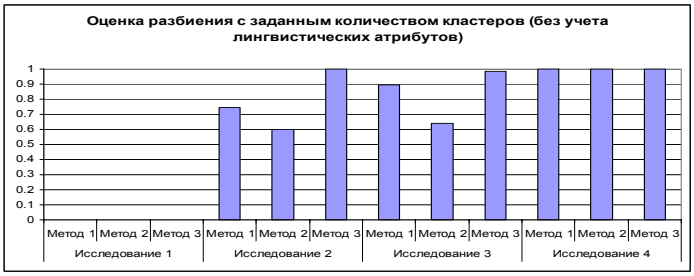



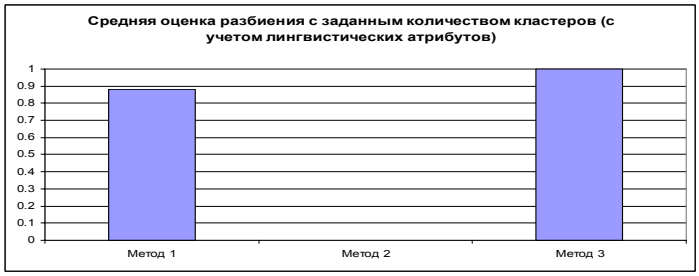



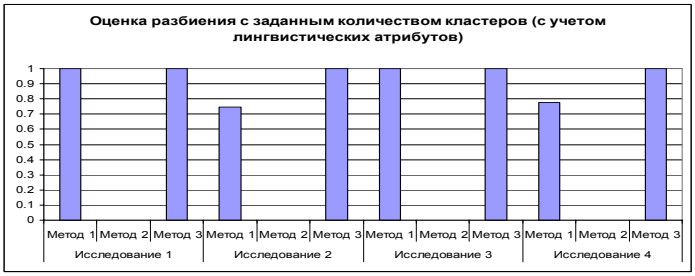

# ОЦЕНКА ЭФФЕКТИВНОСТИ МЕТОДИКИ




# СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МЕТОДИК

- Исследование 1: выделение секторов инвестирования на основе анализа показателей финансовых инструментов;

- Исследование 2: выделение групп клиентов на основе данных об оборотах за период, частоты проведения операций, количестве финансовых инструментов, группы используемых финансовых инструментов;

- Исследование 3: выявление категорий финансовых инструментов для оценки эффективности операций;

- Исследование 4: выделение классов автомобилей на основе данных о максимальной скорости, цвете кузова, воздушном сопротивлении, массе.

Метод 1 – карта Кохонена, Метод 2 – алгоритм к – средних, Метод 3 – ADAKL

## Общая оценка разбиений

| Показатель / Метод | | Количество полученных кластеров | Оценка разбиения | Оценка разбиения с заданным количеством кластеров (без учета лингвистических атрибутов) | Оценка разбиения с заданным количеством кластеров (с учетом лингвистических атрибутов) |
|---|---|---|---|---|---|
| 1 | Метод 1 | 30 | 1.0000 | - | 1.0000 |
| | Метод 2 | - | - | - | - |
| | Метод 3 | 30 | 1.0000 | - | 1.0000 |
| 2 | Метод 1 | 12 | 0.1667 | 0.7467 | 0.7467 |
| | Метод 2 | - | - | 0.6000 | - |
| | Метод 3 | 3 | 1.0000 | 1.0000 | 1.0000 |
| 3 | Метод 1 | 15 | 1.0000 | 0.8953 | 1.0000 |
| | Метод 2 | - | - | 0.6388 | - |
| | Метод 3 | 10 | 1.0000 | 0.9857 | 1.0000 |
| 4 | Метод 1 | 9 | 0.1111 | 1.0000 | 0.7778 |
| | Метод 2 | - | - | 1.0000 | - |
| | Метод 3 | 3 | 1.0000 | 1.0000 | 1.0000 |

## Средняя оценка разбиений

| Оценка / Метод | Оценка разбиения | Оценка разбиения с заданным количеством кластеров (без учета лингвистических атрибутов) | Оценка разбиения с заданным количеством кластеров (с учетом лингвистических атрибутов) | Итоговая оценка |
|---|---|---|---|---|
| Метод 1 | 0.5695 | 0.8807 | 0.8811 | 0.7771 |
| Метод 2 | - | 0.7463 | - | 0.2488 |
| Метод 3 | 1.0000 | 0.9952 | 1.0000 | 0.9984 |

Оценка разбиения выполняется на основе показателей выполненной кластеризации: $O = \frac{r}{n} \times \begin{cases} q/k, q \le k \\ k/q, q > k \end{cases}$,

где

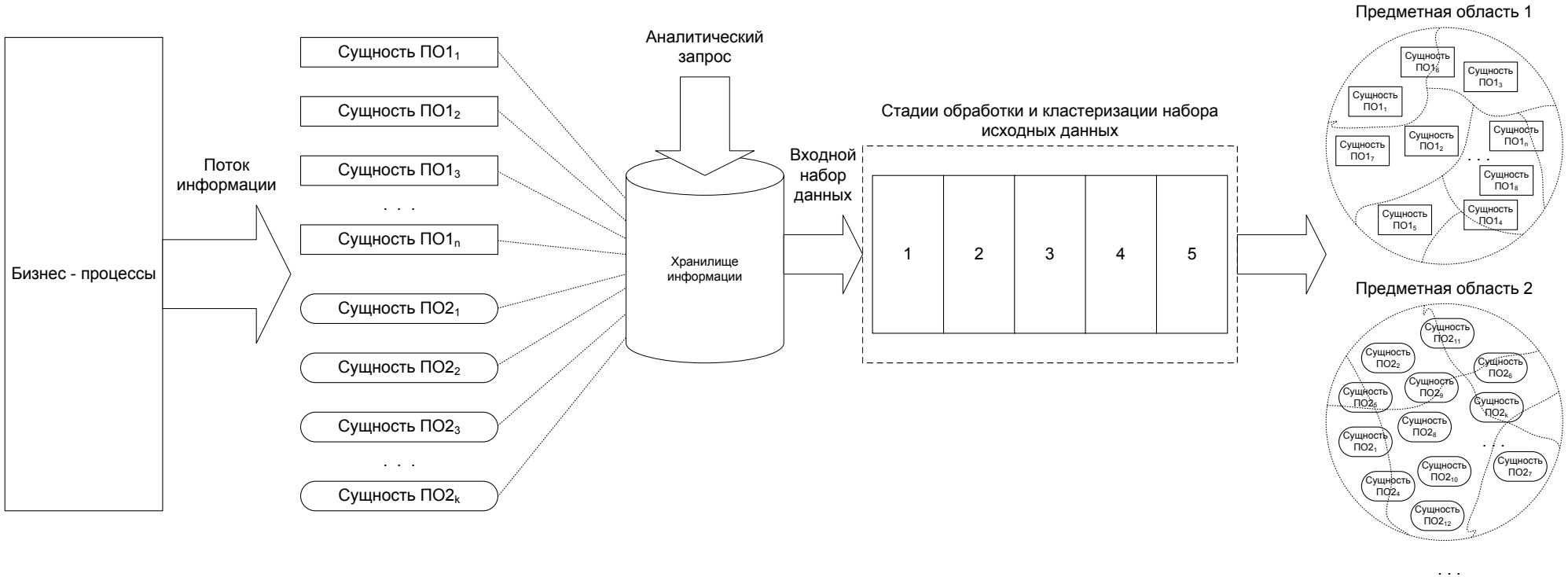
$q$ – количество кластеров по итогам кластеризации;

$r$ – количество элементов, правильно распределенных по соответствующим кластерам;

$k$ – исходное количество кластеров;

$n$ – количество объектов кластеризации.

# Укрупненная схема обработки потока информации




# Укрупненная схема архитектуры программного комплекса