УДК 007.681.5:519.714:519.766

МАТЕМАТИЧЕСКИЙ МЕТОД СК-АНАЛИЗА В СВЕТЕ ИДЕЙ ИНТЕРВАЛЬНОЙ БУТСТРЕПНОЙ РОБАСТНОЙ СТАТИСТИКИ ОБЪЕКТОВ НЕЧИСЛОВОЙ ПРИРОДЫ

Луценко Е.В. – д. э. н., профессор
Кубанский государственный аграрный университет

Интервальные оценки сводят анализ чисел к анализу фактов и позволяют обрабатывать количественные величины как нечисловые данные. Однако это ограничивает возможности обработки количественных величин методами обработки нечисловых данных. В математической модели СК-анализа, основанной на системной теории информации, наоборот, качественным, нечисловым данным приписываются количественные величины. Это позволяет использовать все возможности количественных методов для исследования нечисловых данных. Таким образом, в СК-анализе числовые и нечисловые данные обрабатываются единообразно на основе одной математической модели как числовые данные. Рассматривается связь метода измерения адекватности модели в СК-анализе с бутстрепными методами. Описывается робастная процедура выявления и устранения артефактов в СК-анализе.

1. Постановка проблемы

Современный этап развития информационных технологий характеризуется быстрым ростом производительности компьютеров и облегчением доступа к ним. С этим связан всевозрастающий интерес к использованию компьютерных технологий для организации мониторинга различных объектов, анализа данных, прогнозирования и управления в различных предметных областях. Исследователи и руководители возлагают определенные надежды на повышение эффективности применения компьютерных технологий.

Однако на пути реализации этих ожиданий имеются определенные сложности, связанные с относительным отставанием в развитии математических методов и реализующего их программного инструментария. И анализ, и прогнозирование, и управление существенным образом основываются на математическом моделировании объектов. Математическое моделирование, в свою очередь, предполагают возможность выполнения всех арифметических операций (сложение, вычитание, умножение и деление) над отображениями объектов в моделях и над их элементами.

В практике интеллектуального анализа данных в экономике, социологии, психологии, педагогике и других предметных областях все чаще встречаются ситуации, когда необходимо в рамках единой математической модели *совместно* обрабатывать числовые и нечисловые данные.

Числовые данные могут быть различной природы, и, соответственно, они измеряются в самых различных единицах измерения. Однако арифметические операции можно выполнять только над числовыми данными, измеряемыми в одних единицах измерения.

Данные нечисловой природы, т.е. различные факты и события, характеризуются тем, что с ними вообще нельзя выполнять арифметические операции.

Соответственно, *возникает потребность в математических методах и программном инструментарии, обеспечивающих совместную сопоставимую обработку разнородных числовых данных и данных нечисловой природы.*

2. Традиционные пути решения проблемы

Для проведения подобных исследований обычно реализуется один из двух вариантов:

- изучается подмножество однородных по своей природе данных, измеряемых в одних единицах измерения;

- перед исследованием данные приводятся к сопоставимому виду, например, широко используются процентные или другие относительные величины, реже – стандартизированные значения.

Первый вариант является не решением проблемы, а лишь ее вынужденным обходом, обусловленным ограничениями реально имеющегося в распоряжении исследователей инструментария.

Второй вариант лишь частично решает проблему – снимает различие в единицах измерения. Однако он не преодолевает принципиального различия между количественными и качественными (нечисловыми) величинами и не позволяет обрабатывать их совместно в рамках единой модели.

В последние годы развивается ряд новых методов статистики, полный обзор которых дан в работах А.И. Орлова [1, 2]. Прежде всего, это интервальная статистика, статистика объектов нечисловой природы, робастные, бутстрепные и непараметрические методы.

В частности, методы интервальной статистики позволяют сводить числовые величины к фактам попадания их значений в определенные интервалы, т.е. к событиям. При этом преодолевается проблема возникновения различий в размерности числовых величин, обеспечивается также обработка числовых величин как событий *совместно* с информацией о других событиях, связанных с объектами нечисловой природы. Таким образом, *интервальные методы сводят обработку числовых величин к методам обработки нечисловой информации* и позволяют обрабатывать их *единообразно по одной методике*. Это, в общем-то, вполне очевидный и естественный ход. Однако достигается этот результат *дорогой ценой:* сведением числовых величин к нечисловым, т.е. преобразованием их к "низменному типу", что приводит к утрате ряда возможностей обработки. Это происходит потому, что для числовых величин существует гораздо больше методов и возможностей обработки, чем для нечисловых.

3. Идея решения проблемы

По нашему мнению, более предпочтительным является противоположный подход, основанный на введении некоторой количественной меры, позволяющей единым и сопоставимым образом описывать как числовые данные различной природы, так и нечисловые величины с использованием всего арсенала возможностей, имеющегося при обработке числовых данных.

Приведем аналогию традиционного и предлагаемого решений проблемы на примере обработки документов текстовых редакторов. Если у нас есть документы стандартов "Документ Word" и "Текст-DOS" и мы хотели бы обрабатывать их все в одном редакторе, то это можно осуществить двумя способами:

- преобразовать все документы Word в "низменный стандарт" "Текст-DOS" (аналог традиционного решения проблемы)

- преобразовать "досовские" документы в формат Word (аналог предлагаемого решения проблемы).

В 1979 году автором разработана [3], а в 1981 году впервые применена [4] математическая модель, обеспечивающая реализацию этой идеи. В последующем этот математический аппарат был развит в ряде работ, основной из которых является [5], была разработана соответствующая ему методика численных расчетов, включающая структуры данных и алгоритмы базовых когнитивных операций, а также создана программная система "Эйдос", реализующая математическую модель и методику численных расчетов [6, 7].

Предложенный метод получил название "Системно-когнитивный анализ" (СК-анализ) [5]. В *СК-анализе числовым величинам, так же как и нечисловым, приписываются сопоставимые в пространстве и во времени,*

а также между собой количественные значения, позволяющие обрабатывать их как числовые. Это осуществляется в два этапа:

- числовые величины преобразуются в нечисловые методом интервалов;

- нечисловым величинам, а также преобразованным числовым приписываются числовые значения.

Второй этап является особенностью СК-анализа.

СК-анализ включает следующие этапы:

1. Когнитивная структуризация, а затем и формализация предметной области.

2. Ввод данных мониторинга в базу прецедентов за период, в течение которого имеется необходимая информация в электронной форме.

3. Синтез семантической информационной модели (СИМ).

4. Оптимизация СИМ.

5. Проверка адекватности СИМ (измерение внутренней и внешней, дифференциальной и интегральной валидности).

6. Анализ СИМ.

7. Решение задач идентификации состояний объекта управления, прогнозирование и поддержка принятия управленческих решений по управлению с применением СИМ.

На первых двух этапах СК-анализа, детально рассмотренных в работе [8], числовые величины сводятся к интервальным оценкам, как и информация об объектах нечисловой природы (фактах, событиях). Эти этапы реализуются также в методах интервальной статистики.

На третьем этапе СК-анализа всем этим величинам по единой методике, основанной на системном обобщении семантической теории информации А. Харкевича, приписываются количественные вели-

чины, с которыми в дальнейшем и производятся все операции моделирования.

4. Математическая модель СК-анализа

4.1. Системное обобщение формулы Хартли

Системное обобщение формулы Хартли для равновероятных состояний объекта управления можно представить в виде:

$$I = \mathrm{Log}_2 W. \quad (1) \qquad I = \mathrm{Log}_2 (C_W^1 + C_W^2 + ... + C_W^M), \quad (4)$$

$$I = \mathrm{Log}_2 W^{\varphi}. \quad (2) \qquad \text{при } M = W: \sum_{m=1}^{M} C_W^m = 2^W - 1. \quad (5)$$

$$\boxed{I = \mathrm{Log}_2 \sum_{m=1}^{M} C_W^m}. \quad (3)$$

$$I = \mathrm{Log}_2 (2^W - 1) \approx W,$$

при $W >> 1$; $I \approx W$ с очень малой и быстро уменьшающейся погрешностью, (6)

где W – количество чистых (классических) состояний системы; φ – коэффициент эмерджентности Хартли (уровень системной организации объекта, имеющего W чистых состояний).

4.2. Гипотеза о Законе возрастания эмерджентности

Исследование математических выражений системной теории информации (7–12) позволило сформулировать *гипотезу* о существовании "Закона возрастания эмерджентности". Суть этой гипотезы состоит в том, что в самих элементах системы содержится сравнительно небольшая доля всей записанной в ней информации, а основной ее объем составляет системная информация подсистем различного уровня иерархии.

Различие между классическим и предложенным системными понятиями информации соответствует различию между понятиями МНОЖЕСТВА И СИСТЕМЫ, на основе которых они сформированы.

$$I = \mathrm{Log}_2 W^{\varphi} = \mathrm{Log}_2 \sum_{m=1}^{M} C_W^m. \quad (7)$$

$$\boxed{\varphi = \frac{\mathrm{Log}_2 \sum_{m=1}^{M} C_W^m}{\mathrm{Log}_2 W}.} \quad (8)$$

$$I(W,M) = \mathrm{Log}_2 W^{\frac{\mathrm{Log}_2 \sum_{m=1}^{M} C_W^m}{\mathrm{Log}_2 W}} \quad (9)$$

$$I(W,M) \approx \mathrm{Log}_2 W^{\frac{W}{\mathrm{Log}_2 W}} = W. \quad (10)$$

$$I_{системная} \approx W - \mathrm{Log}_2 W. \quad (11)$$

$$I(W,M) = \mathrm{Log}_2 W + \mathrm{Log}_2 W^{\varphi - 1}. \quad (12)$$

Математическая формулировка

$$\varphi = \frac{\mathrm{Log}_2 \sum_{m=1}^{M} C_W^m}{\mathrm{Log}_2 W} \approx \frac{W}{\mathrm{Log}_2 W},$$

$$\boxed{I_{системная} \approx W - \mathrm{Log}_2 W}.$$

Интерпретация

4.3. Системное обобщение формулы Харкевича

Ниже приведен вывод системного обобщения формулы Харкевича, а именно:

– классическая формула Харкевича через вероятности перехода системы в целевое состояние при условии сообщения ей определенной информации и без использования (13);

– выражение классической формулы Харкевича через частоты (14, 15);

– вывод коэффициента эмерджентности Харкевича на основе принципа соответствия с выражением Хартли в детерминистском случае (16–19);

– вывод системного обобщения формулы Харкевича;

– окончательное выражение для системного обобщения формулы Харкевича (21).

Классическая формула Харкевича

$$I_{ij} = \mathrm{Log}_2 \frac{P_{ij}}{P_j}\ , \qquad (13)$$

где P_{ij} – вероятность перехода объекта управления в j-е состояние в условиях действия i-го фактора; P_j – вероятность самопроизвольного перехода объекта управления в j-е состояние, т.е. в условиях отсутствия действия i-го фактора или в среднем.

Известно, что *корреляция не является мерой причинно-следственных связей*. Если значение корреляции между действием некоторого фактора и переходом объекта управления в определенное состояние высокое, то это не значит, что данный фактор является причиной этого перехода. Для того чтобы по корреляции можно было судить о наличии причинно-следственной связи, необходимо сравнить исследуемую группу с *контрольной группой*, в которой данный фактор не действовал.

Высокая вероятность перехода объекта управления в определенное состояние, так же как и высокая корреляция, в условиях действия некоторого фактора сама по себе не говорит о наличии причинно-следственной связи между ними, т.е. о том, что данный фактор обусловил переход объекта в это состояние. Это связано с тем, что вероятность перехода объекта в это состояние может быть вообще очень высокой и независимо от действия фактора. Поэтому в качестве меры силы причинной обусловленности определенного состояния объекта действием некоторого фактора Харкевич предложил логарифм ***отношения*** вероятностей перехода объекта в это состояние в условиях действия фактора и при его отсутствии или в среднем (13).

Таким образом, семантическая мера информации Харкевича является мерой наличия причинно-следственных связей между факторами и состояниями объекта управления.

Выражение классической формулы Харкевича через частоты фактов

$$P_{ij} = \frac{N_{ij}}{N_i};\ P_i = \frac{N_i}{N};\ P_j = \frac{N_j}{N};$$
$$\text{где } N_i = \sum_{j=1}^{W} N_{ij};\ N_j = \sum_{i=1}^{M} N_{ij};\ N = \sum_{i=1}^{W}\sum_{j=1}^{M} N_{ij}. \qquad (14)$$

$$I_{ij} = \mathrm{Log}_2 \frac{N_{ij} N}{N_i N_j}. \qquad (15)$$

Вывод коэффициента эмерджентности Харкевича на основе принципа соответствия с выражением Хартли в детерминистском случае

Однако мера Харкевича (13), в отличие от меры Шеннона, не удовлетворяет принципу соответствия с мерой Хартли, т.е. не переходит в меру Хартли в детерминистском случае, когда каждому будущему состоянию объекта управления соответствует единственный уникальный фактор и между факторами и состояниями имеется взаимно однозначное соответствие (17).

$$I_{ij} = \mathrm{Log}_2 \left(\frac{N_{ij} N}{N_i N_j} \right)^{\Psi} \qquad (16) \qquad \forall N_{ij} = N_i = N_j = 1. \qquad (17)$$

Откуда

$$I_{ij} = \mathrm{Log}_2 N^{\Psi} = \mathrm{Log}_2 W^{\varphi}. \qquad (18) \qquad \boxed{\Psi = \frac{\mathrm{Log}_2 W^{\varphi}}{\mathrm{Log}_2 N}}. \qquad (19)$$

Вывод системного обобщения формулы Харкевича

$$\Psi = \frac{\mathrm{Log}_2 W^{\frac{\mathrm{Log}_2 \sum_{m=1}^{M} C_W^m}{\mathrm{Log}_2 W}}}{\mathrm{Log}_2 N}. \quad (20)$$

$$I_{ij} = \mathrm{Log}_2\left(\frac{N_{ij}N}{N_i N_j}\right)^{\Psi} = \mathrm{Log}_2\left(\frac{N_{ij}N}{N_i N_j}\right)^{\frac{\mathrm{Log}_2 W^{\varphi}}{\mathrm{Log}_2 N}} =$$

$$= \frac{\mathrm{Log}_2 W^{\varphi}}{\mathrm{Log}_2 N}\left(\mathrm{Log}_2\left(\frac{N_{ij}}{N_i N_j}\right) + \mathrm{Log}_2 N\right) =$$

$$= \mathrm{Log}_2\left(\frac{N_{ij}}{N_i N_j}\right)^{\frac{\mathrm{Log}_2 W^{\varphi}}{\mathrm{Log}_2 N}} + \mathrm{Log}_2 W^{\varphi}.$$

Окончательное выражение для системного обобщения формулы Харкевича

$$\boxed{I_{ij} = \mathrm{Log}_2\left(\frac{N_{ij}}{N_i N_j}\right)^{\frac{\mathrm{Log}_2 W^{\varphi}}{\mathrm{Log}_2 N}} + \mathrm{Log}_2 W^{\varphi}}. \quad (21)$$

4.4. Связь системной теории информации (СТИ) с теорией Хартли – Найквиста – Больцмана и теорией Шеннона

Связь между выражениями для плотности информации в теориях Хартли, Шеннона и СТИ показаны на рисунке 1.

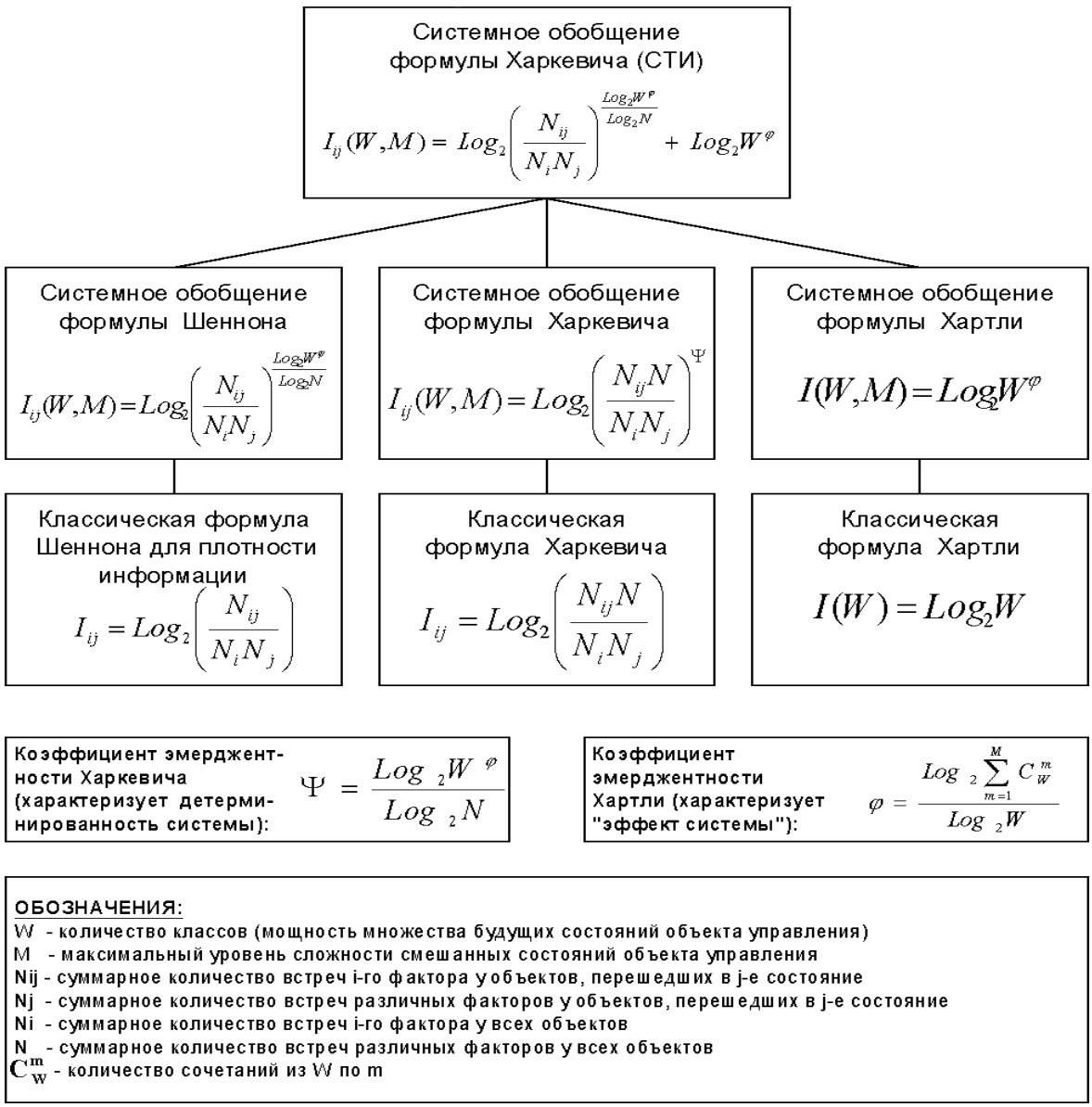

Рис. 1. Связь между выражениями для плотности информации в теориях Хартли, Шеннона и СТИ

4.5. Интерпретация коэффициентов эмерджентности СТИ

Интерпретация коэффициентов эмерджентности, предложенных в рамках системной теории информации, приведена на рисунке 2.

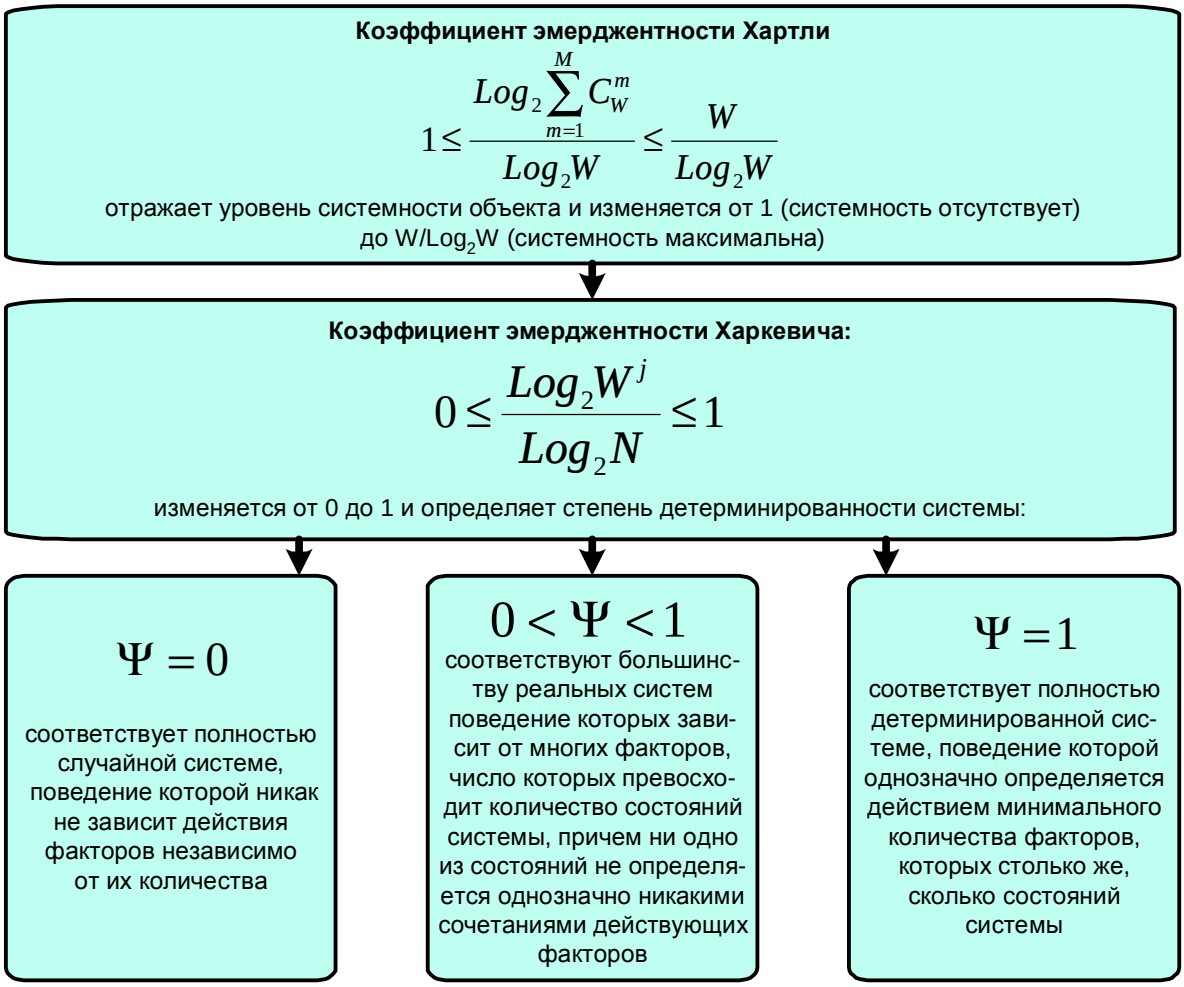

Рис. 2. Интерпретация коэффициентов эмерджентности СТИ

Коэффициент эмерджентности Хартли φ (4) представляет собой относительное превышение количества информации о системе при учете системных эффектов (смешанных состояний, иерархической структуры ее подсистем и т.п.) над количеством информации без учета системности, т.е. этот коэффициент является аналитическим выражением для *уровня системности объекта*.

Коэффициент эмерджентности Харкевича Ψ изменяется от 0 до 1 и определяет степень детерминированности системы.

Таким образом, в предложенном системном обобщении формулы Харкевича (21) впервые непосредственно в аналитическом выражении для самого понятия "Информация" отражены такие фундаментальные свойства

систем, как "Уровень системности" и "Степень детерминированности" системы.

4.6. Матрица абсолютных частот

Основной формой первичного обобщения эмпирической информации в модели является матрица абсолютных частот (табл. 1). В этой матрице строки соответствуют факторам, столбцы – будущим целевым и нежелательным состояниям объекта управления, а на их пересечении приведено количество наблюдений фактов (по данным обучающей выборки), когда действовал некоторый *i*-й фактор и объект управления перешел в некоторое *j*-е состояние.

Таблица 1. МАТРИЦА АБСОЛЮТНЫХ ЧАСТОТ

Атрибуты		Классы - будущие состояния объекта управления					Сумма
		Целевые состояния		Нежелательные состояния			
		***	j	***	l	***	
Факторы характеризующие текущее и прошлые состояния объекта управления, в т.ч. его рефлексивность	***						
	i		N_{ij}		N_{il}		$N_i = \sum_{j=1}^{m} N_{ij}$
	***						
Управляющие факторы системы управления	i		N_{ij}		N_{il}		$N_i = \sum_{j=1}^{m} N_{ij}$
	***						
Факторы характеризующие прошлые текущее и прогнозируемые состояния окружающей среды	k		N_{kj}		N_{kl}		$N_k = \sum_{j=1}^{m} N_{kj}$
	***						
Сумма			$N_j = \sum_{i=1}^{M} N_{ij}$		$N_l = \sum_{i=1}^{M} N_{il}$		$N = \sum_{i=1}^{M} \sum_{j=1}^{m} N_{ij}$

где

N_{ij} – количество встреч i-го признака у объектов j-го класса по данным обучающей выборки

4.7. Матрица информативностей

Непосредственно на основе матрицы абсолютных частот с использованием системного обобщения формулы Харкевича (21) рассчитывается матрица информативностей (табл. 2).

Таблица 2. МАТРИЦА ИНФОРМАТИВНОСТЕЙ

Атрибут		Классы - будущие состояния объекта управления					Средняя детерминирующая мощность i-го вектора
		Целевые состояния		Нецелевые состояния			
		j		l			
Факторы характеризующие текущее и прошлые состояния объекта управления, в т.ч. его рефлексивность	r	$I_{ij} = \Psi \cdot \log_2 \frac{N_{ij} \cdot \sum_{j=1}^{m}\sum_{i=1}^{n} N_{ij}}{\sum_{j=1}^{m} N_{ij} \cdot \sum_{i=1}^{n} N_{ij}}$		$I_{il} = \Psi \cdot \log_2 \frac{N_{il} \cdot \sum_{l=1}^{k}\sum_{i=1}^{n} N_{il}}{\sum_{l=1}^{k} N_{il} \cdot \sum_{i=1}^{n} N_{il}}$			$\sigma_i = \sqrt{\frac{1}{W-1}\sum_{j=1}^{W}(I_{ij} - \bar{I}_i)^2}$
Управляющие факторы системы управления	i	$I_{ij} = \Psi \cdot \log_2 \frac{N_{ij} \cdot \sum_{j=1}^{m}\sum_{i=1}^{n} N_{ij}}{\sum_{j=1}^{m} N_{ij} \cdot \sum_{i=1}^{n} N_{ij}}$		$I_{il} = \Psi \cdot \log_2 \frac{N_{il} \cdot \sum_{l=1}^{k}\sum_{i=1}^{n} N_{il}}{\sum_{l=1}^{k} N_{il} \cdot \sum_{i=1}^{n} N_{il}}$			$\sigma_i = \sqrt{\frac{1}{W-1}\sum_{j=1}^{W}(I_{ij} - \bar{I}_i)^2}$
Факторы характеризующие прошлые текущее и прогнозируемые состояния окружающей среды	k	$I_{kj} = \Psi \cdot \log_2 \frac{N_{kj} \cdot \sum_{j=1}^{m}\sum_{i=1}^{n} N_{ij}}{\sum_{j=1}^{m} N_{kj} \cdot \sum_{i=1}^{n} N_{ij}}$		$I_{kl} = \Psi \cdot \log_2 \frac{N_{kl} \cdot \sum_{l=1}^{k}\sum_{i=1}^{n} N_{il}}{\sum_{l=1}^{k} N_{kl} \cdot \sum_{i=1}^{n} N_{il}}$			$\sigma_k = \sqrt{\frac{1}{W-1}\sum_{j=1}^{W}(I_{kj} - \bar{I}_k)^2}$
Средняя детерминированность будущих состояний АОУ		$\sigma_j = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(I_{ij} - \bar{I}_j)^2}$		$\sigma_l = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(I_{il} - \bar{I}_l)^2}$			$H = \sqrt{\frac{1}{(W-M)\cdot M}\sum_{l=1}^{k}\sum_{i=1}^{n}(I_{il} - \bar{I}_l)^2}$

$\bar{I}_j = \frac{1}{M}\sum_{i=1}^{n} I_{ij}$ – среднее значение коэффициентов вектора класса, M – количество факторов

$\bar{I}_l = \frac{1}{W}\sum_{r=1}^{m} I_{rl}$ – среднее значение коэффициентов вектора фактора, W – количество классов (будущих состояний АОУ).

$\bar{I} = \frac{1}{W \cdot M}\sum_{l=1}^{k}\sum_{i=1}^{n} I_{il}$ – средняя информативность по матрице информативностей

$\varphi = \frac{\log_2 \sum_{j=1}^{m} C_j}{\log_2 M}$ – коэффициент энтропийности Харкевича.

$\psi = \frac{\log_2 W^*}{\log_2 W}$ – коэффициент энтропийности Харкевича.

H – мера уровня определенности плотностной области в рамках СТИ

Матрица информативностей является универсальной формой представления **смысла** эмпирических данных в единстве их дискретного и интегрального представления (причины – последствия, факторы – результирующие состояния, признаки – обобщенные образы классов, образное – логическое, дискретное – интегральное).

Весовые коэффициенты матрицы информативностей непосредственно определяют, какое количество информации I_{ij} система управления получает о наступлении события: "объект управления перейдет в *j*-е состояние" из сообщения: "на объект управления действует *i*-й фактор".

Когда количество информации $I_{ij}>0$, то *i*-й фактор способствует переходу объекта управления в *j*-е состояние, если $I_{ij}<0$ – препятствует этому переходу, $I_{ij}=0$ – никак не влияет на это.

Таким образом, предлагаемая семантическая информационная модель позволяет непосредственно на основе эмпирических данных и независимо от предметной области ***рассчитать, какое количество информации содержится в любом событии о любом другом событии.***

Этот вывод является ключевым для данной статьи, т.к. конкретно показывает возможность числовой обработки в СК-анализе как числовой, так и нечисловой информации.

Матрица информативностей является также обобщенной (неклассической) таблицей решений, в которой входы (факторы) и выходы (будущие состояния объекта управления) связаны друг с другом не с помощью классических (Аристотелевских) импликаций, принимающих только значения: "Истина" и "Ложь", а **различными значениями истинности, выраженными в битах** и принимающими значения от положительного теоретически максимально возможного до теоретически неограниченного отрицательного.

4.8. Неметрический интегральный критерий сходства, основанный на лемме Неймана – Пирсона

В выражениях (22–24) приведен неметрический интегральный критерий сходства, основанный на фундаментальной лемме Неймана – Пирсона и обеспечивающий идентификацию и прогнозирование в предложенных **неортонормированных** семантических пространствах с **финит-**

ной метрикой, ***в которых в качестве координат векторов будущих состояний объекта управления и факторов выступает количество информации, рассчитанное в соответствии с системной теорией информации (21),*** а не Булевы координаты или частоты, как обычно.

$$I_j = \mathrm{f}(\vec{I}_{ij}), \qquad (22) \qquad I_j = (\vec{I}_{ij}, \vec{L}_i) \qquad (23)$$

или в координатной форме

$$I_j = \sum_{i=1}^{M} I_{ij} L_i, \qquad (24) \qquad j^* = \arg\max_{j \in J}((\vec{I}_{ij}, \vec{L}_i)), \qquad (25)$$

где $\vec{I}_{ij} = \{I_{ij}\}$ – вектор j-го состояния объекта управления; $\vec{L}_i = \{L_i\}$ – вектор состояния предметной области, включающий все виды факторов, характеризующих объект управления, возможные управляющие воздействия и окружающую среду (массив-локатор), т.е.

$$\vec{L}_i = \begin{cases} 1, \textit{ если } i-\textit{й фактор действует}; \\ \alpha_i, \textit{ где } 0 < \alpha_i < 1, \textit{ если } i-\textit{й фактор действует с истинностью } \alpha_i. \\ 0, \textit{ если } i-\textit{й фактор не действует}. \end{cases}$$

$$I_j = \frac{1}{\sigma_j \sigma_l M} \sum_{i=1}^{M} \left(I_{ij} - \bar{I}_j\right)\left(L_i - \bar{L}\right), \qquad (26) \qquad I_{ij} \to \frac{I_{ij} - \bar{I}_j}{\sigma_j}, \quad L_i \to \frac{L_i}{\bar{L}} \qquad (27)$$

где $\bar{I}_j$ – средняя информативность по вектору класса; $\bar{L}$ – среднее по вектору идентифицируемой ситуации (объекта); σ_j – среднеквадратичное отклонение информативностей вектора класса; σ_l – среднеквадратичное отклонение по вектору распознаваемого объекта.

4.9. Связь системной меры целесообразности информации с критерием c^2

В (28–33) показана связь системной меры целесообразности информации с известным критерием χ^2, а также предложен новый критерий уровня системности предметной области, являющийся нормированным объемом семантического пространства (34, 35).

$$\chi^2 = \sum_{j=1}^{W}\sum_{i=1}^{M}\frac{(N_{ij}-t)^2}{t}, \qquad (28) \qquad t = \frac{N_i N_j}{N}, \qquad (29)$$

где N_{ij} – фактическое количество встреч i-го признака у объектов j-го класса; t – ожидаемое количество встреч i-го признака у объектов j-го класса.

$$I_{ij} = \mathrm{Log}_2\left(\frac{N_{ij}N}{N_i N_j}\right)^{\Psi}. \qquad (30) \qquad I_{ij} = \mathrm{Log}_2\left(\frac{N_{ij}}{t}\right)^{\Psi}. \qquad (31)$$

$$I_{ij} = \Psi(\mathrm{Log}_2\, N_{ij} - \mathrm{Log}_2\, t). \qquad (32)$$

$$\begin{cases} если\, N_{ij} < t,\, то\, \chi_{ij} > 0, & I_{ij} < 0 \\ если\, N_{ij} = t,\, то\, \chi_{ij} = 0, & I_{ij} = 0 \\ если\, N_{ij} > t,\, то\, \chi_{ij} > 0, & I_{ij} > 0 \end{cases}. \qquad (33)$$

$$H = \sqrt[2]{\frac{1}{(W\,M-1)}\sum_{j=1}^{W}\sum_{i=1}^{M}\left(I_{ij}-\bar{I}\right)^2}. \qquad (34) \qquad \bar{I} = \frac{1}{W\,M}\sum_{j=1}^{W}\sum_{i=1}^{M} I_{ij}. \qquad (35)$$

В качестве более точного критерия уровня системности модели предлагается в данной статье объем неортонормированного семантического пространства, рассчитанный как объем многомерного параллелепипеда, ребрами которого являются оси семантического пространства. Однако для этой меры сложнее в общем виде записать аналитическое выражение, и

для ее вычисления могут быть использованы численные методы с использованием многомерного обобщения смешанного произведения векторов.

Абстрагирование (ортонормирование) значительно сокращает размерность семантического пространства без существенного уменьшения его объема.

4.10. Оценка адекватности семантической информационной модели в СК-анализе и бутстрепные методы

Под адекватностью модели СК-анализа понимается ее внутренняя и внешняя дифференциальная и интегральная валидность. Понятие валидности является уточнением понятия адекватности, для которого определены процедуры количественного измерения, т.е. валидность – это количественная адекватность. Это понятие количественно отражает способность модели давать правильные результаты идентификации, прогнозирования и способность вырабатывать правильные рекомендации по управлению.

Дадим определения следующим понятиям:

- внутренняя валидность – валидность модели, измеренная после ее синтеза путем идентификации объектов обучающей выборки;

- внешняя валидность – валидность модели, измеренная после ее синтеза путем идентификации объектов, не входящих в обучающую выборку;

- дифференциальная валидность модели – достоверность идентификации объектов в разрезе по классам;

- интегральная валидность – средневзвешенная дифференциальная валидность.

Возможны все сочетания: внутренняя дифференциальная валидность, внешняя интегральная валидность и т.д.

Основная идея бутстрепа по Б. Эфрону [9] состоит в том, что методом Монте-Карло (статистических испытаний) многократно извлекаются

выборки из эмпирического распределения. Эти выборки, естественно, являются вариантами исходной, напоминают ее.

Эта идея позволяет сконструировать алгоритм измерения адекватности модели, состоящий из двух этапов:

1. Синтез модели на одном случайном подмножестве обучающей выборки.

2. Измерение валидности модели на оставшемся подмножестве обучающей выборки, не использованном для синтеза модели.

Поскольку оба случайных подмножества имеют переменный состав по объектам обучающей выборки, то подобная процедура должна повторяться много раз, после чего могут быть рассчитаны статистические характеристики адекватности модели, например, такие как:

– средняя внешняя валидность;

– среднеквадратичное отклонение текущей внешней валидности от средней и др.

Достоинство бутстрепного подхода к оценке адекватности модели состоит в том, что он позволяет измерить внешнюю валидность по уже имеющейся выборке и изучить статистические характеристики адекватности модели при изменении объема и состава выборки.

4.11. Непараметричность модели. Робастные процедуры и фильтры для исключения артефактов

Предложенная семантическая информационная модель является ***непараметрической***, т.к. базируется на системной теории информации [5], которая не предполагает нормальности распределений исследуемой выборки.

Под робастными понимаются процедуры, обеспечивающие устойчивую работу модели на исходных данных, зашумленных артефактами, т.е.

выпадающих из общих статистических закономерностей, которым подчиняется исследуемая выборка.

Критерий выявления артефактов, реализованный в СК-анализе, основан на том, что при увеличении объема статистики частоты значимых атрибутов растут, как правило, пропорционально объему выборки, а частоты артефактов так и остаются чрезвычайно малыми, близкими к единице. Таким образом, выявление артефактов возможно только при достаточно большой статистике, т.к. в противном случае недостаточно информации о поведении частот атрибутов с увеличением объема выборки.

В модели реализована такая процедура удаления наиболее вероятных артефактов, и она, как показывает опыт, существенно повышает качество (адекватность) модели.

5. Методика численных расчетов СК-анализа

5.1. Детальный список БКОСА и их алгоритмов

Детальный список базовых когнитивных операций системного анализа, которым соответствуют 24 алгоритма, здесь привести нет возможности из-за их объемности (табл. 3). Они представлены в полном объеме в работе [5].

Таблица 3. ДЕТАЛЬНЫЙ СПИСОК БАЗОВЫХ КОГНИТИВНЫХ ОПЕРАЦИЙ СИСТЕМНОГО АНАЛИЗА (БКОСА)

алго-	по схе- ме	БКОСА	Наименование БКОСА	Полное наименование базовых когнитивных операций системного анализа (БКОСА)
	1.1	1	Присвоение имен	Присвоение имен классам (интенсиональная, интегральная репрезентация)

	1.2			Присвоение имен атрибутам (экстенсиональная, дискретная репрезентация)
1	2.1.	2	Восприятие	Восприятие и запоминание исходной обучающей информации
2	2.2.			Репрезентация. Сопоставление индивидуального опыта с коллективным (общественным)
3	3.1.1	3	Обобщение (синтез, индукция)	Накопление первичных данных
4	3.1.2			Исключение артефактов
5	3.1.3			Расчет истинности смысловых связей между предпосылками и результатами (обобщенных таблиц решений)
6	3.2			Определение значимости шкал и градаций факторов, уровней Мерлина
7	3.3			Определение значимости шкал и градаций классов, уровней Мерлина
8	4.1	4	Абстрагирование	Абстрагирование факторов (снижение размерности семантического пространства факторов)
9	4.2			Абстрагирование классов (снижение размерности семантического пространства классов)

10	5	5	Оценка адекватности	Оценка адекватности информационной модели предметной области
11	7	6	Сравнение, идентификация и прогнозирование	Сравнение, идентификация и прогнозирование. Распознавание состояний конкретных объектов (объектный анализ)
12	9.1	7	Анализ, дедукция и абдукция	Анализ, дедукция и абдукция классов (семантический анализ обобщенных образов классов, решение обратной задачи прогнозирования)
13	9.2			Анализ, дедукция и абдукция факторов (семантический анализ факторов)
14	10.1.1	8	Классификация и генерация конструктов	Классификация обобщенных образов классов
15	10.1.2			Формирование бинарных конструктов классов
16	10.1.3			Визуализация семантических сетей классов
17	10.2.1			Классификация факторов
18	10.2.2			Формирование бинарных конструктов факторов
19	10.2.3			Визуализация семантических сетей факторов
20	10.3.1	9	Содержательное	Содержательное сравнение классов

21	10.3.2		Сравнение	Расчет и отображение многомногозначных когнитивных диаграмм, в т.ч. диаграмм Мерлина
22	10.4.1			Содержательное сравнение факторов
23	10.4.2			Расчет и отображение многомногозначных когнитивных диаграмм, в т.ч. инвертированных диаграмм Мерлина
24	11	10	Планирование и управление	Многовариантное планирование и принятие решения о применении системы управляющих факторов

5.2. Иерархическая структура данных семантической информационной модели СК-анализа

На рисунке 3 приведена в обобщенном виде иерархическая структура баз данных семантической информационной модели системно-когнитивного анализа. На этой схеме базы данных обозначены **прямоугольниками**, а базовые когнитивные операции системного анализа, преобразующие одну базу в другую, – **стрелками** с надписями. Имеются также базовые когнитивные операции, формирующие выходные графические формы. Из этой схемы видно, что одни базовые когнитивные операции готовят данные для других операций, относящихся к более высоким уровням иерархии системы процессов познания. Этим определяется возможная последовательность выполнения базовых когнитивных операций.

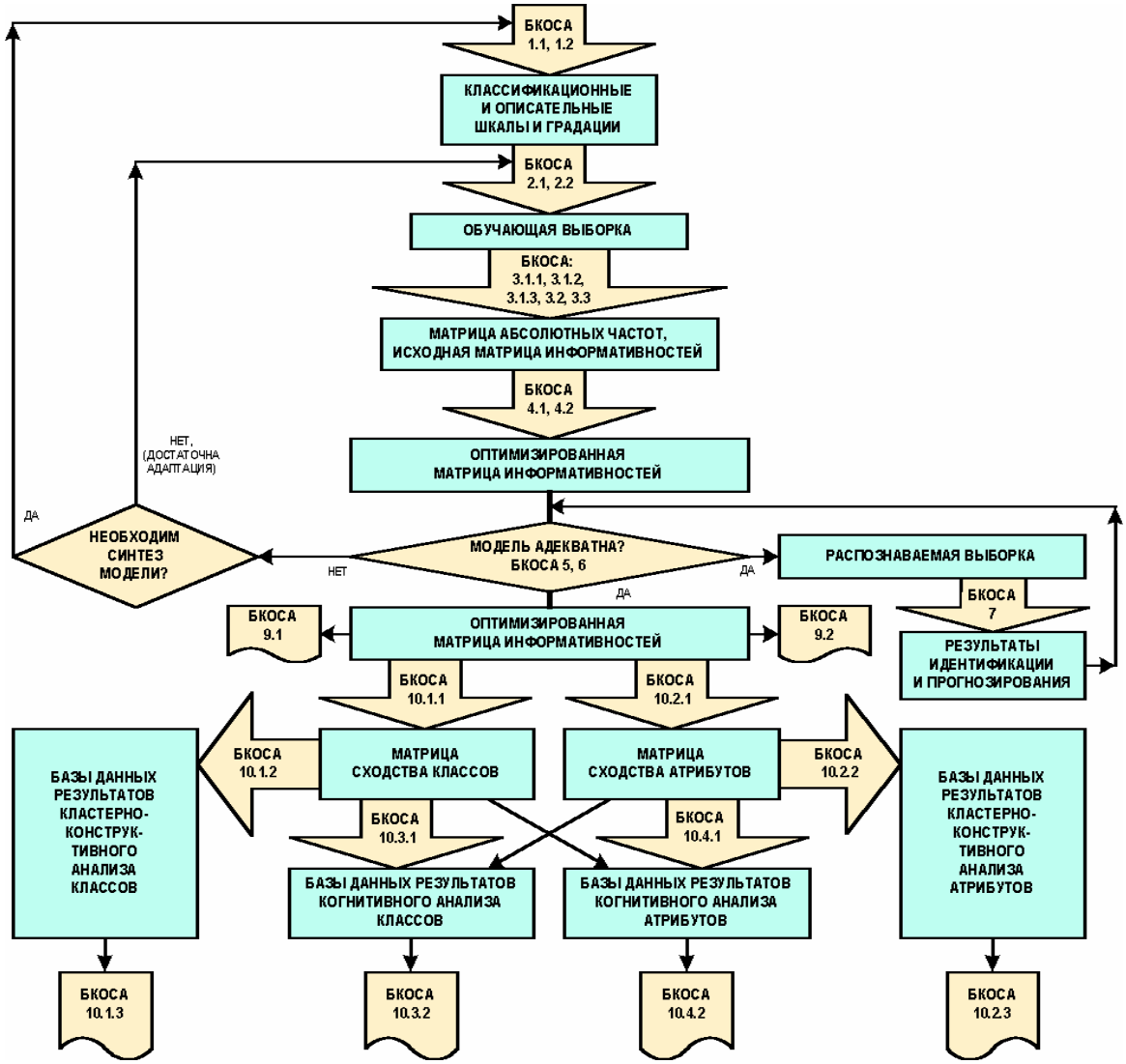

Рис. 3. Иерархическая структура баз данных семантической информационной модели

6. Специальный программный инструментарий СК-анализа – система "Эйдос"

В таблице 4 показана обобщенная схема когнитивной аналитической системы "Эйдос", которая реализует математическую модель и численный метод системно-когнитивного анализа, являющийся его инструментарием.

В состав данной системы входят 7 подсистем.

Первые 3 подсистемы являются **инструментальными**, т.е. позволяют осуществлять **синтез и адаптацию** модели.

Остальные 4 подсистемы обеспечивают идентификацию, прогнозирование и кластерно-конструктивный анализ модели, в т.ч. верификацию модели и выработку управляющих воздействий.

Система "Эйдос" является довольно большой системой: распечатка ее исходных текстов 6-м шрифтом составляет около 800 листов, она генерирует 53 графические формы (двумерные и трехмерные) и 50 текстовых форм. На данную систему и системы окружения получено 8 свидетельств РосПатента РФ.

Таблица 4. ОБОБЩЕННАЯ СТРУКТУРА УНИВЕРСАЛЬНОЙ КОГНИТИВНОЙ АНАЛИТИЧЕСКОЙ СИСТЕМЫ "ЭЙДОС"

№	Подсистема	Режим	Функция	Операция
1	Словари	Классификационные шкалы и градации		
		Описательные шкалы и градации	Наименования шкал	
			Наименования градаций	
		Градации описательных шкал (признаки)		
		Иерархические уровни организации систем	Уровни классов	
			Уровни признаков	
		Почтовая служба по нормативной информации	Обмен по классам	
			Обмен по признакам	
		Печать анкеты		
2	Обучение	Ввод–корректировка обучающей выборки		
		Управление составом обучающей выборки	Параметрическое задание объектов для обработки	
			Статистическая характеристика, ручной ремонт	
			Автоматический ремонт обучающей выборки	
		Пакетное обучение системы распознавания	Накопление абсолютных частот	
			Исключение артефактов (робастная процедура)	
			Расчет информативностей признаков	
			Расчет условных процентных распределений	
			Автоматическое выполнение режимов 1–2–3–4	
			Измерение сходимости и устойчивости модели	Сходимость и устойчивость информационной модели
				Зависимость валидности модели от объема обучающей выборки
		Почтовая служба по обучающей информации		
3	Оптимизация	Формирование ортонормированного базиса классов		
		Исключение признаков с низкой селективной силой		
		Удаление классов и признаков, по которым недостаточно данных		
4	Распознавание	Ввод–корректировка распознаваемой выборки		
		Пакетное распознавание		
		Вывод результатов распознавания	Разрез: один объект – много классов	
			Разрез: один класс – много объектов	
		Почтовая служба по распознаваемой выборке		
5	Типология	Типологический анализ классов распознавания	Информационные (ранговые) портреты (классов)	
			Кластерный и конструктивный анализ классов	Расчет матрицы сходства образов классов
				Генерация кластеров и конструктов классов
				Просмотр и печать кластеров и конструктов
				Автоматическое выполнение режимов: 1,2,3
				Вывод 2d семантических сетей классов
			Когнитивные диаграммы классов	
		Типологический анализ первичных признаков	Информационные (ранговые) портреты признаков	
			Кластерный и конструктивный анализ признаков	Расчет матрицы сходства образов признаков
				Генерация кластеров и конструктов признаков
				Просмотр и печать кластеров и конструктов
				Автоматическое выполнение режимов: 1,2,3
				Вывод 2d семантических сетей признаков
			Когнитивные диаграммы признаков	
6	Анализ	Оценка достоверности заполнения объектов		
		Измерение интегральной и дифференциальной валидности системы распознавания		
		Измерение независимости классов и признаков		
		Просмотр профилей классов и признаков		
		Графическое отображение нелокальных нейронов		
		Отображение Паретто-подмножеств нейронной сети		
7	Сервис	Генерация (сброс) баз данных	Все базы данных	
			Нормативная информация	Всех баз данных
				БД классов
				БД первичных признаков
				БД обобщенных признаков
			Обучающая выборка	
			Распознаваемая выборка	
			Базы данных статистики	
		Переиндексация всех баз данных		
		Печать БД абсолютных частот		
		Печать БД условных процентных распределений		
		Печать БД информативностей		
		Интеллектуальная дескрипторная информационно–поисковая система		

7. Выводы

Интервальные оценки сводят анализ чисел к анализу фактов и позволяют обрабатывать количественные величины как нечисловые данные. Это ограничивает возможности обработки количественных величин методами

обработки нечисловых данных. В математической модели СК-анализа, основанной на системной теории информации, наоборот, качественным, нечисловым данным сопоставляются количественные величины. Это позволяет использовать все возможности количественных методов для исследования нечисловых данных.

Таким образом, в *СК-анализе числовые и нечисловые данные обрабатываются единообразно на основе единой математической модели как числовые данные.*

Рассматривается связь метода измерения адекватности модели в СК-анализе с бутстрепными методами.

Описывается робастная процедура выявления и устранения артефактов в СК-анализе.

Список литературы

1. Орлов А.И. Надежность и контроль качества. 1991. № 8. С. 3–8.

2. Орлов А.И. Современная прикладная статистика. http://www.mira-ech.com.ua/linkst_art.htm.

3. Луценко Е.В. Автоматизированная система распознавания образов: математическая модель и опыт применения // В.И. Вернадский и современность (к 130-летию со дня рождения): Сборник. – Краснодар: КНА, 1993. – С. 37–42.

4. Луценко Е.В. Теоретические основы и технология адаптивного семантического анализа в поддержке принятия решений (на примере универсальной автоматизированной системы распознавания образов "ЭЙДОС-5.1"). – Краснодар: КЮИ МВД РФ, 1996. – 280 с.

5. Луценко Е.В. Автоматизированный системно-когнитивный анализ в управлении активными объектами (системная теория информации и ее применение в исследовании экономических, социально-психологических, технологических и организационно-технических систем): Монография (научное издание). – Краснодар: КубГАУ, 2002. – 605 с.

6. Пат. № 940217. РФ. Универсальная автоматизированная система распознавания образов "ЭЙДОС" / Е.В. Луценко (Россия); Заяв. № 940103. Опубл. 11.05.94. – 50 с.

7. Пат. № 2003610986 РФ. Универсальная когнитивная аналитическая система "ЭЙДОС" / Е.В. Луценко (Россия); Заяв. № 2003610510 РФ. Опубл. от 22.04.2003. – 50 с.

8. Луценко Е.В. Типовая методика и инструментарий когнитивной структуризации и формализации задач в СК-анализе // Научный журнал КубГАУ. – 2004.– № 1 (3). –18 с. http://ej.kubagro.ru

9. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – М.: Финансы и статистика, 1988. – 263 с.