УДК 303.732.4

СЕМАНТИЧЕСКАЯ ИНФОРМАЦИОННАЯ МОДЕЛЬ СК-АНАЛИЗА

Луценко Евгений Вениаминович
д. э. н., к. т. н., профессор

Кубанский государственный аграрный университет, Краснодар, Россия

В статье кратко описывается семантическая информационная модель системно-когнитивного анализа (СК-анализ), вводится универсальная информационная мера силы и направления влияния значений факторов, независимая от их природы и единиц измерения на поведение объекта управления (основанная на лемме Неймана – Пирсона), а также неметрический интегральный критерий сходства между образами конкретных объектов и обобщенными образами классов, образами классов и образами значений факторов. Идентификация и прогнозирование рассматривается как разложение образа конкретного объекта в ряд по обобщенным образам классов (объектный анализ), что предлагается рассматривать как возможный вариант решения на практике 13-й проблемы Гильберта.

Ключевые слова: СЕМАНТИЧЕСКАЯ ИНФОРМАЦИОННАЯ МОДЕЛЬ, СИСТЕМНО-КОГНИТИВНЫЙ АНАЛИЗ (СК-АНАЛИЗ), ИНФОРМАЦИОННАЯ МЕРА СИЛЫ И НАПРАВЛЕНИЯ ВЛИЯНИЯ ЗНАЧЕНИЙ ФАКТОРОВ, ПОВЕДЕНИЕ ОБЪЕКТА УПРАВЛЕНИЯ, ЛЕММА НЕЙМАНА – ПИРСОНА, НЕМЕТРИЧЕСКИЙ ИНТЕГРАЛЬНЫЙ КРИТЕРИЙ СХОДСТВА.

UDC 303.732.4

SEMANTIC INFORMATION MODEL OF SC-ANALYSIS

Lutsenko Evgeny Veniaminovich
Dr. Sci. Econ., Cand. Tech. Sci., professor

Kuban State Agrarian University, Krasnodar, Russia

Semantic informational model of systemic-cognitive analysis (SC-analysis) is described in this article in brief, universal information measure of force and direction influence of factors values, independent from their nature and measurement units on the action of an object’s management (based on the lemma Neiman-Pierson), and nonmetric integral criterion of similarity between images of concrete objects and generic mages of classes, images of classes and images of factors values are introduced as well. Identification and prognosis are considered as a decomposition of an image of a concrete object in a row by generic images classes (object analysis), it is offered to consider as a possible variant of decision of the thirteen problem of Gilbert in practice.

Key words: SEMANTIC INFORMATION MODEL, SYSTEMIC-COGNITIVE ANALYSIS (SC-ANALYSIS), INFORMATION MEASURE OF FORCE AND DIRECTION INFLUENCE OF FACTORS VALUES, ACTION OF MANAGEMENT OBJECT, LEMMA OF NEIMAN AND PIERSON, NONMETRIC INTEGRAL CRITERION OF SIMILARITY.

Количество информации в индивидуальных событиях и лемма Неймана – Пирсона

В классическом анализе Шеннона идет речь лишь о передаче символов по одному информационному каналу от одного источника к одному приемнику. Его интересует, прежде всего, передача самого сообщения.

В данном исследовании ставится другая задача: идентифицировать информационный источник по сообщению от него. Поэтому метод Шеннона был обобщен путем учета в математической модели возможности существования многих источников информации, о которых к приемнику по зашумленному каналу связи приходят не отдельные символы-признаки, а сообщения, состоящие из последовательностей символов (признаков) любой длины.

Следовательно, ставится задача идентификации информационного источника по сообщению от него, полученному приемником по зашумленному каналу. Метод, являющийся обобщением метода К. Шеннона, позволяет применить классическую теорию информации для построения моделей систем распознавания образов и принятия решений, ориентированных на применение для синтеза адаптивных АСУ сложными объектами.

Для решения поставленной задачи необходимо вычислять не средние информационные характеристики, как в теории Шеннона, а количество информации, содержащееся в конкретном i*–м признаке (символе) о том, что он пришел от данного* j*–го источника информации. Это позволит определить суммарное количество информации в сообщении о каждом информационном источнике, что дает интегральный критерий для идентификации или прогнозирования состояния АОУ.*

Логично предположить, что среднее количество информации, содержащейся в системе признаков о системе классов

$$I(Y,X)=\sum_{j=1}^{W}\sum_{i=1}^{M}p_{ij}Log_2\frac{p_{ij}}{p_i p_j} \tag{1}$$

является усреднением (с учетом условной вероятности наблюдения) "индивидуальных количеств информации", которые содержатся в конкретных признаках о конкретных классах (источниках), т.е.:

$$i(x_j,y_i)=Log_2\frac{p_{ij}}{p_i p_j}. \tag{2}$$

Выражение (2) определяет так называемую "плотность информации", т.е. количество информации, которое содержится в одном отдельно взятом факте наблюдения i-го символа (признака) на приемнике о том, что этот символ (признак) послан j-м источником.

Если в сообщении содержится M символов, то суммарное количество информации о принадлежности данного сообщения j-у информационному источнику (классу) составляет:

$$i(x_j)=\sum_{i=1}^{M}Log_2\frac{p_{ij}}{p_i p_j}. \tag{3}$$

Необходимо отметить, что применение сложения в выражении (3) является вполне корректным и оправданным, так как информация с самого начала вводилась как аддитивная величина, для которой операция сложения является корректной.

Преобразуем выражение (3) к виду, более удобному для практического применения (численных расчетов). Для этого выразим вероятности встреч признаков через частоты их наблюдения:

$$P_{ij} = \frac{1}{N_{ij}};\ P_i = \frac{1}{N_i};\ P_j = \frac{1}{N_j}. \qquad (4)$$

Подставив (4) в (3), получим:

$$i(x_j) = \sum_{i=1}^{M} Log_2 \frac{N_{ij}}{N_i N_j}. \qquad (5)$$

Если ранжировать классы в порядке убывания суммарного количества информации о принадлежности к ним, содержащейся в данном сообщении (т.е. описании объекта), и выбирать первый из них, т.е. тот, о котором в сообщении содержится наибольшее суммарное количество информации, то мы получим обоснованную статистическую процедуру, основанную на классической теории информации, оптимальность которой доказывается в фундаментальной лемме Неймана – Пирсона [1].

Таким образом, распознавание образов предполагает принятие решения о принадлежности объекта или его состояния к определенному классу. Если до распознавания существовала неопределенность в вопросе о том, к какому классу относятся распознаваемый объект или его состояние, то в результате распознавания эта неопределенность уменьшается, в том числе может быть равна нулю. Понятие "информация" может быть определено, как "количественная мера степени снятия неопределенности". Количество информации является мерой соответствия распознаваемого объекта (его состояния) обобщенному образу класса.

Количество информации имеет ряд вполне определенных свойств. Эти свойства позволяют ввести понятие "количество информации в индивидуальных событиях", которое является весьма перспективным для применения в системах распознавания образов и поддержки принятия решений.

Математическая модель метода распознавания образов и принятия решений, основанного на системной теории информации

Решение задачи 1: "Синтез семантической информационной модели активного объекта управления"

Исходные данные для выявления взаимосвязей между факторами и состояниями объекта управления предлагается представить в виде корреляционной матрицы – матрицы абсолютных частот (таблица 1).

В этой матрице в качестве классов (столбцов) приняты будущие состояния объекта управления как целевые, так и нежелательные, а в качестве атрибутов (строк) – факторы, которые разделены на три основных группы, математически обрабатываемые единообразно: факторы, характеризующие текущие и прошлые состояния объекта управления; управляющие факторы системы управления; факторы, характеризующие прошлые, текущие и прогнозируемые состояния окружающей среды. Отметим, что форма таблицы 1 является универсальной для представления и обобщения ***фактов – эмпирических данных*** в единстве их дискретного и интегрального представления (причины – следствия, факторы – результирующие состояния, признаки – обобщенные образы классов, образное – логическое и т.п.).

Управляющие факторы объединяются в группы, внутри каждой из которых они альтернативны (несовместны), а между которыми – нет (совместны). В этом случае внутри каждой группы выбирают одно из доступных управляющих воздействий с максимальным влиянием. Варианты содержательной информационной модели без учета прошлых состояний объекта управления и с их учетом аналогичны, соответственно, простым и составным цепям Маркова, автоматам без памяти и с памятью.

Таблица 1 – МАТРИЦА АБСОЛЮТНЫХ ЧАСТОТ

Атрибуты		Классы – будущие состояния объекта управления					Сумма
		Целевые состояния		Нежелательные состояния			
		...	j	...	j	...	
Факторы, характеризующие текущее и прошлые состояния объекта управления и его окружения	...		N_{ij}		N_{ij}		$N_i=\sum N_{ij}$
Управляющие факторы системы управления	i		N_{ij}		N_{ij}		$N_i=\sum_{j=1}^{W} N_{ij}$
Факторы, характеризующие прошлые, текущие и прогнозируемые состояния окружающей среды	k		N_{ij}		N_{ij}		$N_i=\sum_{j=1}^{W} N_{ij}$
Сумма			$N_j=\sum_{i=1}^{M} N_{ij}$		$N_j=\sum_{i=1}^{M} N_{ij}$		$N=\sum_{i=1}^{W}\sum_{j=1}^{M} N_{ij}$

где:

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

В качестве количественной меры влияния факторов, предложено использовать обобщенную формулу А. Харкевича (6), полученную на основе предложенной эмерджентной (системной) теории информации:

$$I_{ij} = Log_2\left(\frac{N_{ij}}{N_i N_j}\right)^{\frac{Log_2 W^j}{Log_2 N}} + Log_2 W^j \qquad (6)$$

При этом по формуле (6) непосредственно из матрицы абсолютных частот (см. таблицу 1) рассчитывают матрицу информативностей (таблица 2), *которая представляет собой основу содержательной информационной модели предметной области*.

Весовые коэффициенты (6) таблицы 2 непосредственно определяют, какое количество информации I_{ij} система управления получает о наступлении события: "активный объект управления перейдет в j-е состояние", из сообщения: "на активный объект управления действует i-й фактор".

Таблица 2 – МАТРИЦА ИНФОРМАТИВНОСТЕЙ

[illegible]		[illegible]		[illegible]	[illegible]
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]
[illegible]		[illegible]		[illegible]	[illegible]

[illegible]

Принципиально важно, что эти весовые коэффициенты не определяются экспертами неформализуемым способом, а *рассчитываются* <u>*непосредственно*</u> *на основе эмпирических данных* и удовлетворяют всем ранее сформулированным требованиям, т.е. являются сопоставимыми, содержательно интерпретируемыми, отражают понятия "достижение цели управления" и "мощность множества будущих состояний объекта управления" и т.д.

В работе [1] показано, что предложенная информационная мера обеспечивает ***сопоставимость*** индивидуальных количеств информации, содержащейся в факторах (***независимо от их природы***) о классах, а также сопоставимость интегральных критериев, рассчитанных для одного объекта и разных классов, для разных объектов и разных классов.

Когда количество информации I_{ij}**>0** – ***i*-й** фактор способствует переходу объекта управления в ***j*-е** состояние, когда I_{ij}**<0** – препятствует этому переходу, когда же I_{ij}**=0** – никак не влияет на это. В векторе ***i*-го** фактора (строка матрицы информативностей) отображается, какое количество информации о переходе объекта управления в каждое из будущих состояний содержится в том факте, что данный фактор действует. В векторе ***j*-го** состояния класса (столбец матрицы информативностей) отображается, какое количество информации о переходе объекта управления в соответствующее состояние содержится в каждом из факторов.

Таким образом, матрица информативностей (см таблицу 2) является обобщенной таблицей решений, в которой входы (факторы) и выходы (будущие состояния АОУ) 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 ("Максимальная степень истинности"), до теоретически неограниченного отрицательного ("Степень ложности").*

Фактически предложенная модель позволяет осуществить *синтез обобщенных таблиц решений* для различных предметных областей непосредственно на основе эмпирических исходных данных и *продуцировать на их основе прямые и обратные правдоподобные (нечеткие) логические рассуждения по неклассическим схемам* ***с различными расчетными значениями истинности****,* являющимися обобщением классических импликаций (таблица 3).

Таблица 3 – ПРЯМЫЕ И ОБРАТНЫЕ ПРАВДОПОДОБНЫЕ ЛОГИЧЕСКИЕ ВЫСКАЗЫВАНИЯ С РАСЧЕТНОЙ (В СООТВЕТСТВИИ С СТИ) СТЕПЕНЬЮ ИСТИННОСТИ ИМПЛИКАЦИЙ

	Прямые высказывания:	Обратные высказывания
1	***если A, то B*** (если действует фактор A, то мы предполагаем с степенью истинности I_{AB}, что АОУ перейдет в состояние B)	***если B, то A*** (если АОУ перешел в состояние B, то мы предполагаем со степенью истинности I_{AB}, что действовал фактор A)
2	***если A_1 и A_2 ... и A_K, то B*** (прогноз влияния системы факторов на поведение АОУ. Степень истинности обобщающей (итоговой) импликации равна алгебраической сумме истинностей составляющих ее элементарных импликаций вида: "если A то B")	***если B, то A_1 и A_2 ... и A_K*** (информационный портрет класса B, т.е. влияние различных факторов A_i на переход АОУ в будущее состояния B, решение обратной задачи прогнозирования, т.е. выработка управления)
3	***если A, то B_1 или B_2 ... или B_N*** (семантический портрет фактора A, т.е. его влияние на переход АОУ в различные состояния)	
4	***если A_1 и A_2 ... и A_K, то B_1 или B_2 ... или B_N*** (прогноз влияния системы факторов на переход АОУ в различные состояния)	

Приведем пример более сложного высказывания, которое может быть рассчитано непосредственно на основе матрицы информативностей, которая, по сути, является обобщенной таблицей решений (см. таблицу 2).

Если A со степенью истинности a(A,B) детерминирует B и если C со степенью истинности a(C,D) детерминирует D, и A совпадает по смыслу с C со степенью истинности a(A,C), то это вносит вклад в совпадение B с D, равный степени истинности a(B,D).

При этом в прямых рассуждениях как предпосылки рассматриваются факторы, а как заключение – будущие состояния АОУ, а в обратных – наоборот: как предпосылки – будущие состояния АОУ, а как заключение – факторы. Степень истинности ***i***-й предпосылки – это просто количество информации I_{ij}, содержащейся в ней о наступлении ***j***-го будущего состояния АОУ. Если предпосылок несколько, то *степень истинности наступления **j**-го состояния АОУ равна суммарному количеству информации, содержащемуся в них об этом*. Количество информации в ***i***-м факторе о наступлении ***j***-го состояния АОУ рассчитывается в соответствии с выражением (6) СТИ.

Прямые правдоподобные логические рассуждения позволяют *прогнозировать* степень достоверности наступления события по действующим факторам, а обратные – по заданному состоянию восстановить степень необходимости и степень нежелательности каждого фактора для наступления этого состояния, т.е. принимать решение по выбору *управляющих воздействий* на АОУ, оптимальных для перевода его в заданное целевое состояние.

Необходимо отметить, что предложенная модель, основывающаяся на теории информации, обеспечивает ***автоматизированное формирования системы нечетких правил по содержимому входных данных,*** *как и комбинация нечеткой логики Заде – Коско с нейронными сетями Кохонена*. Принципиально важно, что качественное изменение модели путем добавления в нее новых классов не уменьшает достоверности распознавания уже сформированных классов. Кроме того, при сравнении распознаваемого объекта с каждым классом учитываются не только признаки, имеющиеся у объекта, но и отсутствующие у него, поэтому *предложенной моделью правильно идентифицируются объекты, признаки которых образуют множества, одно из которых является подмножеством другого* (как и в Неокогнитроне К. Фукушимы) [1].

Данная модель позволяет прогнозировать поведение АОУ при воздействии на него не только одного, но и целой ***системы*** факторов:

$$I_j = f(\vec{I}_{ij}). \tag{7}$$

В теории принятия решений скалярная функция I_j векторного аргумента называется ***интегральным критерием***. Основная проблема состоит в выборе или нахождении такого аналитического *вида функции* интегрального критерия, который обеспечил бы эффективное решение сформулированной выше задачи АСУ.

Учитывая, что частные критерии (6) имеют смысл количества информации, а информация, по определению, является аддитивной функцией, предлагается ввести *интегральный критерий* как аддитивную функцию от частных критериев в виде:

$$I_j = (\vec{I}_{ij}, \vec{L}_i). \tag{8}$$

В выражении (8) в круглых скобках обозначено скалярное произведение. В координатной форме это выражение имеет вид:

$$I_j = \sum_{i=1}^{A} I_{ij} L_i, \tag{9}$$

где

$\vec{\mathrm{I}}_{\mathrm{ij}} = \{\mathrm{I}_{\mathrm{ij}}\}$ – вектор *j*-го состояния объекта управления;

$\vec{\mathrm{L}}_{\mathrm{i}} = \{\mathrm{L}_{\mathrm{i}}\}$ – вектор состояния предметной области, включающий все виды факторов, характеризующих объект управления, возможные управляющие воздействия и окружающую среду (массив – локатор), т.е.

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

В реализованной модели значения координат вектора состояния предметной области принимались равными либо 1 (фактор действует), либо 0 (фактор не действует).

Таким образом, <u>*интегральный критерий*</u> *представляет собой суммарное количество информации, содержащееся в системе факторов различной природы (т.е. факторах, характеризующих объект управления, управляющее воздействие и окружающую среду) о переходе активного объекта управления в будущее (в т.ч. целевое или нежелательное) состояние.*

В многокритериальной постановке <u>задача идентификации, распознавания или прогнозирования</u> состояния объекта управления, при оказании на него заданного многофакторного управляющего воздействия I_j, сводится к максимизации интегрального критерия:

$$j^* = \arg\max_{j \in J}((\vec{I}_{ij}, \vec{L}_i)), \tag{10}$$

т.е. к выбору такого состояния объекта управления, для которого интегральный критерий ***максимален****.*

<u>Задача принятия решения</u> о выборе наиболее эффективного управляющего воздействия является *обратной задачей* по отношению к задаче максимизации интегрального критерия (идентификации и прогнозирования), т.е. вместо того, чтобы по набору факторов прогнозировать будущее состояние активного объекта управления (АОУ), наоборот, по заданному (целевому) состоянию АОУ определяется такой набор факторов, который с наибольшей эффективностью перевел бы объект управления в это состояние.

Предлагается еще одно обобщение этой фундаментальной леммы, основанное на косвенном учете корреляций между информативностями в векторе состояний при использовании средних по векторам. Соответственно, вместо простой суммы количеств информации предлагается использовать корреляцию между векторами состояния и объекта управления, которая количественно измеряет степень сходства этих векторов:

$$I_j = \frac{1}{\sigma_j \sigma_l A} \sum_{i=1}^{M} \left(I_{ij} - \bar{I}_j \right) \left(L_i - \bar{L} \right), \qquad (11)$$

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.

Выражение (11) получается непосредственно из (9) после замены координат перемножаемых векторов их стандартизированными значениями:

$$I_{ij} \rightarrow \frac{I_{ij} - \bar{I}_j}{\sigma_j}, \quad L_i \rightarrow \frac{L_i - \bar{L}}{\sigma_l}.$$

Необходимо отметить, что выражение для интегрального критерия сходства (11) по своей математической форме является корреляцией двух векторов. Это означает, что если вектора являются суммой двух сигналов: полезного и белого шума, то при расчете интегрального критерия *белый* шум практически не будет играть никакой роли, т.е. его корреляция с самим собой равна нулю по определению. Это означает, что ***выбранный интегральный критерий сходства является высокоэффективным средством подавления белого шума и выделения полезной информации из шума,*** который неизбежно присутствует в эмпирических данных.

Важно также отметить неметрическую природу интегрального критерия, благодаря чему является корректным его применение при неортонормированном семантическом информационном пространстве, т.е. в общем случае.

Результат прогнозирования поведения объекта управления, описанного данной системой факторов, представляет собой список его возможных будущих состояний, в котором они расположены в порядке убывания суммарного количества информации о переходе объекта управления в каждое из них.

Зависимость уровня системности модели от ее ортонормирован-

ности

Рассмотрим выражение (12):

$$\varphi = \frac{\mathrm{Log}_2 \sum_{m=1}^{M} C_W^m}{\mathrm{Log}_2 W}. \qquad (12)$$

При выполнении операции ортонормирования по классам из модели последовательно удаляют те из них, которые наиболее сильно корреляционно связаны друг с другом. В результате, в модели остаются классы, практически не коррелирующие, т.е. ортонормированные. Поэтому предположим, что *в результате ортонормирования правила запрета на образование подсистем классов становятся более жесткими, и уровень системности модели уменьшается*.

Зависимость степени детерминированности модели от ее ортонормированности

Рассмотрим выражение (13):

$$\Psi = \frac{\mathrm{Log}_2 W^{\varphi}}{\mathrm{Log}_2 N}. \qquad (13)$$

Так как каждый класс, как правило, описан более чем одним признаком, то *при ортонормировании классов* и удалении некоторых из них из модели суммарное количество признаков N будет уменьшаться быстрее, чем количество классов W, поэтому *степень детерминированности будет возрастать*.

При ортонормировании атрибутов числитель выражения (13) не изменяется, а знаменатель уменьшается, поэтому и в этом случае *степень детерминированности возрастает*.

Таким образом, ***ортонормирование модели приводит к увеличению степени ее детерминированности***. По этой причине предлагается считать "детерминированностью" и "системностью" модели не их значения в текущем состоянии модели, а тот предел, к которому стремятся эти величины при корректном ортонормировании модели *при достижении ею максимума адекватности*.

Взаимосвязь системной меры целесообразности информации со статистикой c^{2} ***и новая мера уровня системности предметной области***

Статистика χ^2 представляет собой сумму вероятностей совместного наблюдения признаков и объектов по всей корреляционной матрице или определенным ее подматрицам (т.е. сумму относительных отклонений частот совместного наблюдения признаков и объектов от среднего):

$$c^2 = \sum_{j=1}^{W} \sum_{i=1}^{M} \frac{(N_{ij} - t)^2}{t}, \qquad (14)$$

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.

$$t = \frac{N_i N_j}{N}. \qquad (15)$$

Отметим, что статистика χ^2 математически связана с количеством информации в системе признаков о классе распознавания в соответствии с системным обобщением формулы Харкевича для плотности информации (6):

$$I_{ij} = Log_2 \left(\frac{N_{ij} N}{N_i N_j} \right)^{\Psi}, \qquad (16)$$

а именно, из (15) и (16) получаем:

$$I_{ij} = Log_2 \left(\frac{N_{ij}}{t} \right)^{\Psi}. \qquad (17)$$

Из (17) очевидно:

$$I_{ij} = \Psi(Log_2 N_{ij} - Log_2 t). \qquad (18)$$

Сравнивая выражения (14) и (18), видим, что числитель в выражении (14) под знаком суммы отличается от выражения (18) только тем, что в выражении (18) вместо значений N_{ij} и t взяты их логарифмы. Так как логарифм является монотонно возрастающей функцией аргумента, то введение логарифма не меняет общего характера поведения функции.

Фактически это означает, что:

$$\begin{cases} \text{если } N_{ij} < t \text{ то } \chi_{ij} > 0,\ I_{ij} < 0 \\ \text{если } N_{ij} = t \text{ то } \chi_{ij} = 0,\ I_{ij} = 0. \\ \text{если } N_{ij} > t \text{ то } \chi_{ij} > 0,\ I_{ij} > 0 \end{cases} \qquad (19)$$

Если фактическая вероятность наблюдения i-го признака при предъявлении объекта j-го класса равна ожидаемой (средней), то наблюдение этого признака **не несет никакой информации о принадлежности объекта к данному классу.** Если она выше средней, то это свидетельствует в пользу того, что предъявлен объект данного класса, если ниже – то другого.

Поэтому наличие статистической связи (информации) между признаками и классами распознавания, т.е. отличие вероятностей их совместных наблюдений от предсказываемого в соответствии со случайным нормальным распределением, приводит к увеличению фактической статистики χ^2 по сравнению с теоретической величиной.

Из этого следует возможность использования в качестве количественной меры степени выраженности закономерностей в предметной области не матрицы абсолютных частот и меры χ^2, а новой меры H, основанной на матрице информативностей и системном обобщении формулы Харкевича для количества информации:

$$H = \sqrt[2]{\frac{1}{(W \cdot M - 1)} \sum_{j=1}^{W} \sum_{i=1}^{M} \left(I_{ij} - \bar{I} \right)^2}\ , \tag{20}$$

где

$\bar{I} = \frac{1}{W \cdot M} \sum_{j=1}^{W} \sum_{i=1}^{M} I_{ij}$ – средняя информативность признаков по матрице информативностей.

Меру H в выражении (20) предлагается назвать обобщенным критерием степени сформированности модели Харкевича.

Значение данной меры показывает среднее отличие количества информации в факторах о будущих состояниях активного объекта управления от среднего количества информации в факторе (которое при больших выборках близко к 0). По своей математической форме эта мера сходна с мерами для значимости (интегральной информативности) факторов и степени сформированности образов классов и коррелирует с объемом неортонормированного семантического информационного пространства классов и семантического информационного пространства атрибутов.

Вышеописанная математическая модель обеспечивает инвариантность результатов ее синтеза относительно следующих параметров обучающей выборки: *суммарное количество и порядок ввода анкет обучающей выборки;* количество анкет обучающей выборки по каждому классу распознавания; суммарное количество признаков во всех анкетах обучающей выборки; суммарное количество признаков по эталонным описаниям различных классов распознавания; количество признаков и их порядок в отдельных анкетах обучающей выборки.

Это обеспечивает высокую степень качества решения задач системой распознавания на неполных и разнородных (в вышеперечисленных аспектах) данных как обучающей, так и распознаваемой выборки, т.е. при таких статистических характеристиках потоков этих данных, которые чаще всего и встречается на практике и которыми невозможно или очень сложно управлять.

Сравнение, идентификация и прогнозирование как разложение векторов объектов в ряд по векторам классов (объектный анализ)

В разделе 3.2.3 [1] были введены неметрические интегральные критерии сходства объекта, описанного массивом-локатором L_i с обобщенными образами классов I_{ij} (выражения 7– 9):

$$I_j = (\vec{I}_{ij}, \vec{L}_i). \qquad (21)$$

В выражении (21) круглыми скобками обозначено скалярное произведение. В координатной форме выражение (21) имеет вид:

$$I_j = \sum_{i=1}^{M} I_{ij} L_i, \qquad (22)$$

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.:

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

Для непрерывного случая выражение (22) принимает вид:

$$I_j = \int_1^M L(i) I_j(i) di. \qquad (23)$$

Таким образом, выражение (23) представляет собой вариант выражения (22) интегрального критерия сходства объекта и класса для непрерывного случая в координатной форме.

Отметим, что коэффициенты ряда Фурье по своей математической форме и смыслу сходны с ненормированными коэффициентами корреляции, т.е. по сути скалярными произведениями для непрерывных функций в координатной форме: выражение (23) между разлагаемой в ряд кривой f(x) и функциями Sin и Cos различных частот и амплитуд на отрезке [–L, L] [1].

Из сравнения выражений (23) и (24) следует вывод о том, что ***процесс идентификации и прогнозирования (распознавания), реализованный в предложенной математической модели, может рассматриваться <u>как разложение вектора-локатора распознаваемого объекта в ряд по векторам информативностей классов распознавания</u> (которые представляют собой произвольные функции, сформированные при синтезе модели на основе эмпирических данных).***

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{npx}{L}\right) + b_n \sin\left(\frac{npx}{L}\right) \right)$$

где :

$$a_0 = \frac{1}{L} \int_{-L}^{L} f(x) dx$$

$$a_n = \frac{1}{L} \int_{-L}^{L} f(x) \cos\left(\frac{npx}{L}\right) dx \quad , \qquad (24)$$

$$b_n = \frac{1}{L} \int_{-L}^{L} f(x) \sin\left(\frac{npx}{L}\right) dx$$

где $n=\{1, 2, 3,\ldots\}$ – натуральное число.

Например, результаты идентификации представим на рисунке.

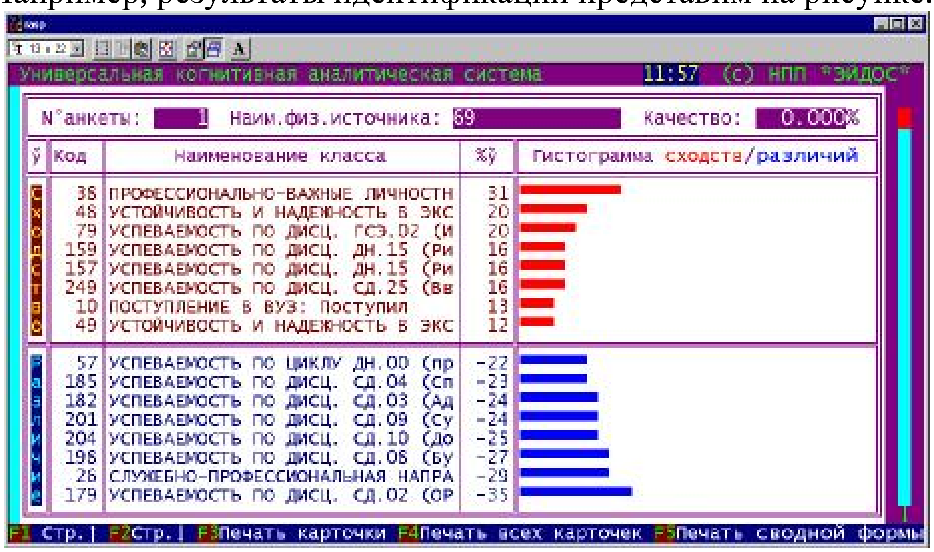

Рисунок – Пример разложения профиля курсанта усл.№69 в ряд по обобщенным образам классов

Продолжая развивать аналогию с разложением в ряд, данный результат идентификации можно представить в векторной аналитической форме:

$$\vec{K}_{усл.№} = 0{,}31 \cdot \vec{I}(38) + 0{,}20 \cdot \vec{I}(48) + 0{,}20 \cdot \vec{I}(79) + 0{,}16 \cdot \vec{I}(159) +$$
$$+ 0{,}16 \cdot \vec{I}(157) + 0{,}16 \cdot \vec{I}(249) + \ldots - 0{,}35 \cdot \vec{I}(179) - 0{,}29 \cdot \vec{I}(26) -$$
$$- 0{,}27 \cdot \vec{I}(198) - 0{,}25 \cdot \vec{I}(204) - 0{,}24 \cdot \vec{I}(201) - 0{,}24 \cdot \vec{I}(182) - \ldots$$

или в координатной форме, более удобной для численных расчетов:

$$\boxed{K(i) = \sum_{j=1}^{W} \left(I(j) \cdot I(i, j) \right)}, \qquad (25)$$

где

I(j) – интегральный критерий сходства массива-локатора, описывающего состояние объекта и ***j***-го класса, рассчитываемый, согласно выражению (26):

$$I_j = \frac{1}{s_j s_l M} \sum_{i=1}^{M} \left(I_{ij} - \bar{I}_j\right)\left(L_i - \bar{L}\right), \tag{26}$$

I(i,j) – вектор обобщенного образа ***j***-го класса, координаты которого рассчитываются в соответствии с системным обобщением формулы Харкевича (27):

$$\mathrm{I_{ij}} = \mathrm{Log_2}\left(\frac{\mathrm{N_{ij}}}{\mathrm{N_i N_j}}\right)^{\frac{\mathrm{Log_2 W^{\varphi}}}{\mathrm{Log_2 N}}} + \mathrm{Log_2 W^{\varphi}}. \tag{27}$$

<u>Примечание:</u> обозначения *I(i,j)* и I_{ij}, и т.п. эквивалентны. Смысл всех переменных, входящих в выражения (27) и (26) раскрыт в разделе 3.1.3 работы [1].

При дальнейшем развитии данной аналогии естественно возникают вопросы: о полноте, избыточности и ортонормированности системы векторов классов как функций, по которым проводится разложение вектора объекта; о сходимости, т.е. вообще возможности и корректности такого разложения.

В общем случае вектор объекта совершенно не обязательно должен разлагаться в ряд по векторам классов таким образом, что сумма ряда во всех точках точно совпадала со значениями исходной функции. Это означает, что система векторов классов может быть ***неполна*** по отношению к профилю распознаваемого объекта, и, тем более, всех возможных объектов.

Предлагается считать не разлагаемые в ряд, т.е. плохо распознаваемые объекты, суперпозицией хорошо распознаваемых объектов ("похожих" на те, которые использовались для формирования обобщенных образов классов), и объектов, которые и не должны распознаваться, так как объекты этого типа <u>не встречались в обучающей выборке и не использовались</u> для формирования обобщенных образов классов, а также не относятся к представляемой обучающей выборкой генеральной совокупности.

Нераспознаваемую компоненту можно рассматривать либо как шум, либо считать ее полезным сигналом, несущим ценную информацию о неисследованных объектах интересующей нас предметной области (в зависимости от целей и тезауруса исследователей). Первый вариант не приводит к осложнениям, так как примененный в математической модели алгоритм сравнения векторов объектов и классов, основанный на вычислении нормированной корреляции Пирсона (сумма произведений), является *весьма устойчивым к наличию белого шума* в идентифицируемом сигнале. Во втором варианте необходимо дообучить систему распознаванию объектов, несущих такую компоненту (в этой возможности и заключается адаптивность модели). Технически этот вопрос решается просто копированием описаний плохо распознавшихся объектов из распознаваемой выборки в обучающую, их идентификацией экспертами и дообучением системы. Кроме того, может быть целесообразным расширить справочник классов распознавания новыми классами, соответствующими этим объектам.

Однако на практике гораздо чаще наблюдается противоположная ситуация (можно даже сказать, что она типична), когда система векторов ***избыточна***, т.е. в системе классов распознавания есть очень похожие классы (между которыми имеет место высокая корреляция, наблюдаемая в режиме: "кластерно-конструктивный анализ"). Практически это означает, что в системе сформировано несколько практически одинаковых образов с разными наименованиями. Для исследователя это само по себе является очень ценной информацией. Однако если исходить только из потребности разложения распознаваемого объекта в ряд по векторам классов (чтобы определить суперпозицией каких образов он является, т.е. "разложить его на компоненты"), то наличие сильно коррелирующих друг с другом векторов представляется неоправданным, так как просто увеличивает размерности данных, внося в них мало нового по существу. Поэтому возникает задача *исключения избыточности системы классов распознавания*, т.е. выбора из всей системы классов распознавания такого минимального их набора, в котором профили классов минимально коррелируют друг с другом, *т.е. ортогональны в фазовом пространстве признаков*. Это условие в теории рядов называется "ортонормируемостью" системы базовых функций, а в факторном анализе связано с идеей выделения "главных компонент".

В предлагаемой математической модели реализованы два варианта выхода из данной ситуации:

1) исключение неформирующихся, расплывчатых классов;

2) объединение почти идентичных по содержанию (дублирующих друг друга) классов.

Однако выбрать нужный вариант и реализовать его, используя соответствующие режимы, пользователь технологии АСК-анализа должен сам. Вся необходимая и достаточная информация для принятия соответствующих решений предоставляется пользователю инструментария АСК-анализа.

Если считать, что функции образов составляют формально-логическую систему, к которой применима теорема Геделя, то можно сформулировать эту теорему для данного случая следующим образом: "Для любой системы базисных функций в принципе всегда может существовать по крайней мере одна такая функция, что она не может быть разложена в ряд по данной системе базисных функций, т.е. ***функция, которая является ортонормированной ко всей системе базисных функций в целом".*** Поэтому для адекватного отражения подобных функций в модели необходимо повышение размерности семантического информационного пространства.

Очевидно, не взаимосвязанными друг с другом могут быть только четко оформленные, детерминистские образы, т.е. образы с высокой степенью редукции ("степень сформированности конструкта"). Поэтому в процессе выявления взаимно-ортогональных базисных образов, в первую очередь, будут выброшены аморфные "расплывчатые" образы, которые связаны практически со всеми остальными образами.

В некоторых случаях результат такого процесса представляет интерес, и это делает оправданным его реализацию. Однако можно предположить, что наличие расплывчатых образов в системе является оправданным, так как в этом случае система образов не будет формальной и подчиняющейся теореме Геделя. Следовательно, система распознавания будет более полна в том смысле, что увеличится вероятность идентификации *любого объекта*, предъявленного ей на распознавание. Конечно, уровень сходства с аморфным образом не может быть столь высоким, как с четко оформленным. Поэтому в этом случае более уместно применить термин "ассоциация" или нечеткая, расплывчатая идентификация, чем "однозначная идентификация".

Итак, можно сделать следующий вывод: допустимость в математической модели СК-анализа не только четко оформленных (детерминистских) образов, но и образов аморфных, нечетких, расплывчатых является важным достоинством данной модели. Это обусловлено тем, что данная модель обеспечивает корректные результаты анализа, идентификации и прогнозирования даже в тех случаях, когда модели идентификации и информационно-поисковые системы детерминистского типа традиционных АСУ практически неработоспособны. В этих условиях данная модель СК-анализа работает как система *ассоциативной (нечеткой) идентификации.*

Таким образом, в предложенной семантической информационной модели при идентификации и прогнозировании, по сути, осуществляется разложение векторов идентифицируемых объектов по векторам классов распознавания, т.е. осуществляется ***"объектный анализ"*** *(по аналогии с спектральным, гармоническим или* ***Фурье-анализом),*** ***что позволяет рассматривать идентифицируемые объекты как суперпозицию обобщенных образов классов различного типа с различными амплитудами (25). При этом вектора обобщенных образов классов, с математической точки зрения, представляют собой произвольные функции и не обязательно образуют полную и не избыточную (ортонормированную) систему функций.***

Для любого объекта всегда существует такая система базисных функций, что вектор объекта может быть представлен в форме линейной суперпозиции (суммы) этих базисных функций с различными амплитудами. Это утверждение, по-видимому, является одним из следствий фундаментальной теоремы А.Н. Колмогорова, доказанной им в 1957 году (О представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения // Докл. АН СССР, Т. 114, С. 953–956, 1957).

Теорема Колмогорова: Любая непрерывная функция от n переменных $F(x_1, x_2, ..., x_n)$ может быть представлена в виде:

$$F(x_1, x_2, ..., x_n) = \sum_{j=1}^{2n+1} \left(g_j \sum_{i=1}^{n} \left(h_{ij}(x_i) \right) \right),$$

где g_j и h_{ij} – непрерывные функции, причем h_{ij} не зависят от функции F.

Эта теорема означает, что для реализации функций многих переменных достаточно операций суммирования и композиции функций одной переменной. Удивительно, что в этом представлении лишь функции g_j зависят от представляемой функции F, а функции h_{ij} универсальны. ***Необходимо отметить, что теорема Колмогорова является обобщением теоремы В.И. Арнольда (1957), которая дает решение 13-й проблемы Гильберта.***

К сожалению, определение вида функций h_{ij} и g_j для данной функции F представляет собой математическую проблему, для которой пока не найдено общего строгого решения.

В данной работе предлагается рассматривать предлагаемую семантическую информационную модель как <u>один из вариантов</u> решения этой проблемы. <u>В этом контексте функция F интерпретируется как образ идентифицируемого объекта, функция h_{ij} – образ j-го класса, а функция g_j – мера сходства образа объекта с образом класса.</u>

Список литературы

1. 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.

2. Луценко Е.В. Интеллектуальные информационные системы: Учебное пособие для студентов специальности "Прикладная информатика (по областям)" и другим экономическим специальностям. 2-е изд., перераб. и доп. – Краснодар: КубГАУ, 2006. – 615 с.

3. Луценко Е.В. Теоретические основы и технология адаптивного семантического анализа в поддержке принятия решений (на примере универсальной автоматизиро-

ванной системы распознавания образов "ЭЙДОС-5.1"). – Краснодар: КЮИ МВД РФ, 1996. – 280 с.

4. Симанков В.С., Луценко Е.В. Адаптивное управление сложными системами на основе теории распознавания образов. Монография (научное издание). – Краснодар: ТУ КубГТУ, 1999. – 318 с.

5. Симанков, В.С. Системный анализ в адаптивном управлении: Монография (научное издание) / В.С. Симанков, Е.В. Луценко, В.Н. Лаптев В.Н..; Под науч. ред. В.С. Симанкова. – Краснодар: ИСТЭК КубГТУ, 2001. – 258 с.

6. Луценко Е.В. АСК-анализ как метод выявления когнитивных функциональных зависимостей в многомерных зашумленных фрагментированных данных // Научный журнал КубГАУ [Электронный ресурс]. – Краснодар: КубГАУ, 2005. – №03(11). – Режим доступа: http://ej.kubagro.ru/2005/03/19/p19.asp.

7. Луценко Е.В. Количественные меры возрастания эмерджентности в процессе эволюции систем (в рамках системной теории информации) // Научный журнал КубГАУ [Электронный ресурс]. – Краснодар: КубГАУ, 2006. – №05(21). – Режим доступа: http://ej.kubagro.ru/2006/05/pdf/31.pdf.

8. Луценко Е.В. Системно-когнитивный анализ как развитие концепции смысла Шенка – Абельсона // Научный журнал КубГАУ [Электронный ресурс]. – Краснодар: КубГАУ, 2004. – №03(5). – Режим доступа: http://ej.kubagro.ru/2004/03/04/p04.asp.

9. Луценко Е.В. Виртуализация общества как основной информационный аспект глобализации (основы информационно-функциональной теории развития техники и информационной теории стоимости) // Научный журнал КубГАУ [Электронный ресурс]. – Краснодар: КубГАУ, 2005. – №01(9). – Режим доступа: http://ej.kubagro.ru/2005/01/02/p02.asp.

10. Луценко Е.В. Критерии реальности и принцип эквивалентности виртуальной и "истинной" реальности // Научный журнал КубГАУ [Электронный ресурс]. – Краснодар: КубГАУ, 2004. – №06(8). – Режим доступа: http://ej.kubagro.ru/2004/06/10/p10.asp.

Примечание:

Для обеспечения доступа читателей к некоторым из этих и другим работам авторов они размещены в Internet по адресам:
http://lc.kubagro.ru/aidos/eidos.htm
http://ej.kubagro.ru/a/viewaut.asp?id=11.