

Entropy Analysis Markup Language – язык метаописаний для систем интеллектуального анализа данных

Аннотация. Статья посвящена разработке языка разметки EAML для формализации описания структуры данных и расчетных модулей энтропийного метода анализа. Определяются основные приемы составления системных профилей, интегрирующих возможности отладки, тестирования и визуализации результатов анализа. Приведен пример аналитического шаблона на языке EAML с заключениями для сложной социально-экономической системы.

Ключевые слова: метаописание, анализ данных, энтропийный метод, многокритериальный выбор, системный аудит, функциональные показатели, сложные системы.

Введение

Одним из наиболее значительных достижений в области информационных систем и искусственного интеллекта стала разработка мощных систем интеллектуального анализа данных.

В качестве аргумента в обоснование необходимости развития и реализации подходов, характерных для анализа данных, выдвигается известный тезис об уникальности, беспрецедентности объектов наблюдения и/или их совокупности и, следовательно, об отсутствии или низкой надежности образцов видов состояния (особенно критических) для выбора модели и управления.

Уникальность объекта(ов) наблюдения рассматривается как обобщение методических трудностей, которые возникают при работе с реальными данными – отсутствие доказательств полноты группы, принадлежности к одной генеральной совокупности, эргодичности, а также различие в природе и размерности показателей, скрытые корреляции и систематические погрешности показателей, вытекающие из методик измерения, малые размеры выборок, асимметрия таблиц, пробелы в данных.

Применение метаописаний в системах анализа данных и системах извлечения данных получило широкое распространение. В работах [1, 2] представлены программные агенты-

скауты, позволяющие на основе метаописания KGL (Knowledge Generation Language), извлекать или синтезировать кортежи предметной информации из «больших» локальных и распределенных баз данных.

Стандартным инструментом для решения задач анализа данных и разработки, решающих правил стали технологии OLAP и Data Mining, а для формализации решающих правил – инструментальные средства и на основе специализированных языков высокого уровня с объектно-ориентированными моделями типа CLIPS и LISP [3].

Существенным недостатком реализуемых таким образом специализированных программных пакетов для анализа данных является их статичность – один или несколько устойчивых тестов (критериев) адаптированных под конкретный набор данных и реализованных в едином алгоритме от ввода данных до визуализации результатов.

В существующих технологиях без внимания остались более гибкие подходы, обеспечивающие интеграцию компонентов аналитических систем общими синтаксическими правилами, например, метаописанием их поведения. Вполне логичным было бы отображать взаимодействие модулей загрузки, подготовки, обработки и анализа данных в виде разветвленной иерархической схемы действий, приводящей к совокупности (комплексу)

аналитических заключений, логически и методически связанных между собой.

Именно такой подход реализован и развивается в РПК «ЭнтропияПлюс» [4]. Разветвленная логика энтропийного метода анализа [5,6], положенного в основу РПК, позволяет в полной мере реализовать и эффективно использовать новый подход к организации аналитических систем на базе языка метаописаний EAML.

1. Задачи исследования

1. Разработать определяющие синтаксические и семантические правила языка метаописаний структур и алгоритмов энтропийного метода анализа (ЭМА) на базе расширяемого языка разметки (XML). Создать структуру системно-аналитического профиля (САП) согласно общему определению типа документа (DTD) с сохранением полиморфизма. Ввести систему "ссылочных множеств", представленных уникальными (id) и ссылочными (parentid) идентификаторами для каждого из методов схемы САП с последовательной структурой размещения методов в схеме, где «расчетный выход» одного модуля будет являться «входом» для другого.

2. Создать определение типа документа (DTD) системно-аналитического профиля и его XML-описание согласно DTD. Структуру EAML-документа представить последовательным набором элементов-тэгов, включаемых в корневые элементы-дескрипторы секций экспертного поведения системы. Такие секции разделить на два типа: секция данных (элементы-тэги <matrix>) и секция решений (элементы-тэги <schemas>). Секция данных должна содержать элементы, определяющие источник информации об анализируемой системе: таблицу объектов и показателей, задающих ее поведение; элемент описывающий способ доступа к этому источнику, служебные тэги, необходимые для работы клиент-серверных составляющих экспертной системы. Секция решений должна состоять из элементов, задающих схему действий экспертной системы. Схема заполняется тэгами, определяющими математические преобразования секции данных.

2. Концепция EAML

Структура САП определяется математической моделью энтропийного метода анализа

[5,6]. Алгоритмы применения его для различных типов данных и математические обоснования приведены в [4].

Далее предлагается сфокусировать свое внимание непосредственно на прикладном аспекте метода и метаописании его математического аппарата в рамках анализа данных, где правомерность всех преобразований вытекает из теоретических основ энтропийного анализа, доказывается и обосновывается в его изложении, поэтому авторы принимают данный подход, как аксиому.

Энтропийный метод анализа разрабатывался с ориентацией на данные сложноструктурированные и неоднородные, содержащие показатели различной природы и размерности, предполагающий возможности построения моделей сколь угодно высокой сложности, но на единой алгоритмической основе. Эта идея определила выбор основных отображений данных, обеспечивающих аддитивность частных (простейших моделей показателей). Тогда более сложные модели, получаемые комбинацией частных, ограничиваются только условиями интерпретируемости, наличием здравого смысла – объективных предпосылок, физических, эмпирических, полуэмпирических моделей показателей.

Иерархическая структура реализации системно-аналитического профиля энтропийного анализа является следствием перечисленных выше свойств выборочных данных, и отражает эвристическую природу поиска в системе, представленной таблицей данных подсистем и связей показателей.

На старте аналитического процесса возможные результаты и выводы неизвестны, подсистемы и связи показателей только предполагаются, а гипотезы о них только оформляются в виде сложных энтропийных моделей.

В энтропийном методе анализа выделяют следующие математические модели: энтропийная модель выборочных данных, логарифмическая модель, обобщенные и комбинированные модели, инвертированные модели и обобщенные элементы [5, 6] (для простоты ссылок, назовем их математическими преобразованиями). Именно правила корректности записи последовательности преобразований в плоскости ЭМА и определяют семантику языка EAML. Синтаксис и вид записи преобразований в виде дерева отражает эвристически структурированный поиск каких-то особенностей данных. В его

структуре можно будет комбинировать различные математические модели ЭМА, что выражает результат поиска в виде гистограмм и фазовых портретов.

Семантика записи последовательности математических моделей ЭМА позволяет выделить каждое преобразование в отдельный элемент (тэг) со своим набором атрибутов, задающих правила для интерпретатора. Целесообразно будет называть каждый элемент согласно его математической модели. Пример одной из базовых моделей ЭМА и их семантические соответствия элементам EAML показаны на Рис. 1.

Логарифмирование и инвертирование накладывает определенные вычислительные ограничения на исходные данные. Для обеспечения универсальности ЭМА был введен вспомогательный класс элементов: процедуры сдвига элементов в каждом столбце матрицы при отрицательном или нулевом значении в ячейки, замена цифры ноль в ячейки на среднее значение в данном столбце, масштабирование

при малых значениях в ячейках матрицы. Пустые ячейки в матрице заменяются на нулевые значения, затем обрабатываются этими же процедурами. Поэтому в семантику описание EAML необходимо было ввести соответствующие элементы. На Рис. 2 показана схема одной из таких процедур в EAML.

Обобщенные и комбинированные модели ЭМА позволяют суммировать элементы матрицы, которые подчиняются переместительному закону (значения в таблице исходных данных с одинаковой фундаментальной характеристикой в одной системе СИ, например столбцы с показателями напряжения (Вольт), добыча (Тонна) и т.п.). Отсюда имеем возможность определить обобщенную по подмножеству показателей энтропийную модель (Рис. 3)

В зависимости от размерности исходных данных показатели, имеющие один и тот же смысл и отражающие одно и то же фундаментальное свойство объекта (например, масса, теплотворность угля), после отображения логарифмической, инвертированной или комбини-

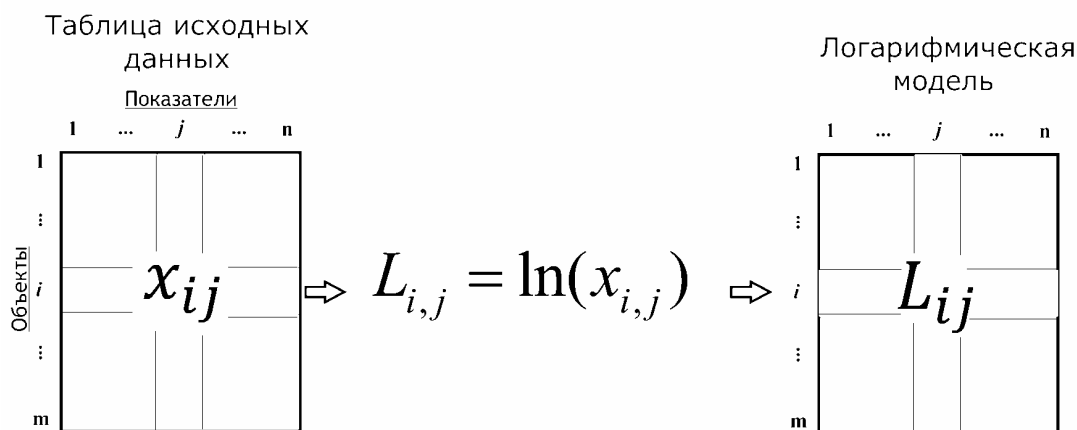


Рис. 1. Схема математической модели ЭМА для логарифмической модели

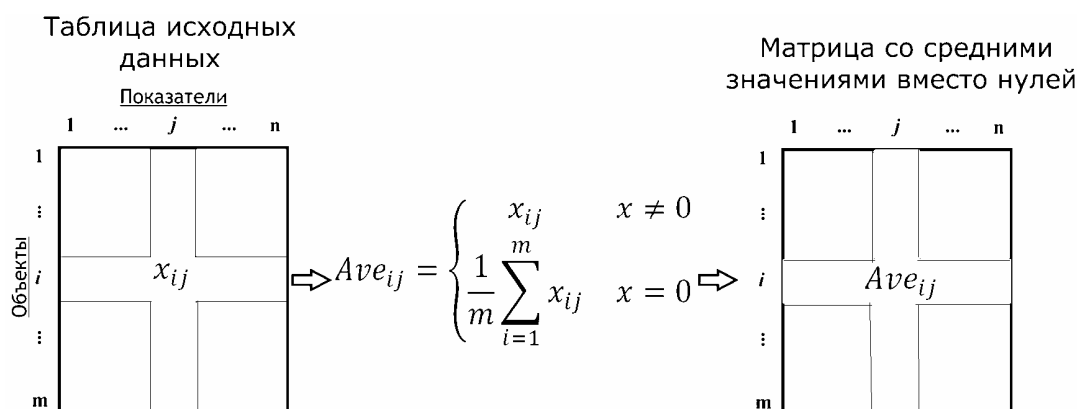


Рис. 2. Схема процедуры сдвига матрицы при отрицательном или нулевом значении

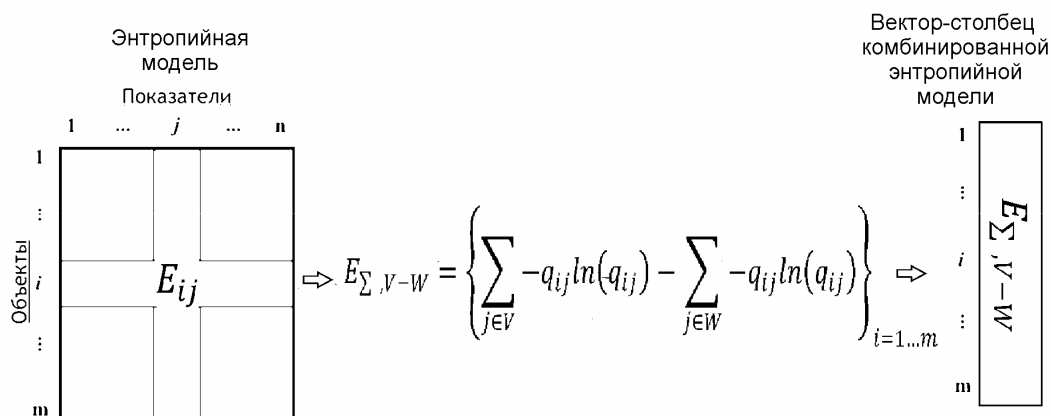


Рис. 3. Схема процедуры комбинирования (по столбцам) энтропийной модели для элемента, где V и W – множества вектор-столбцов объектов с положительными и отрицательными характеристиками

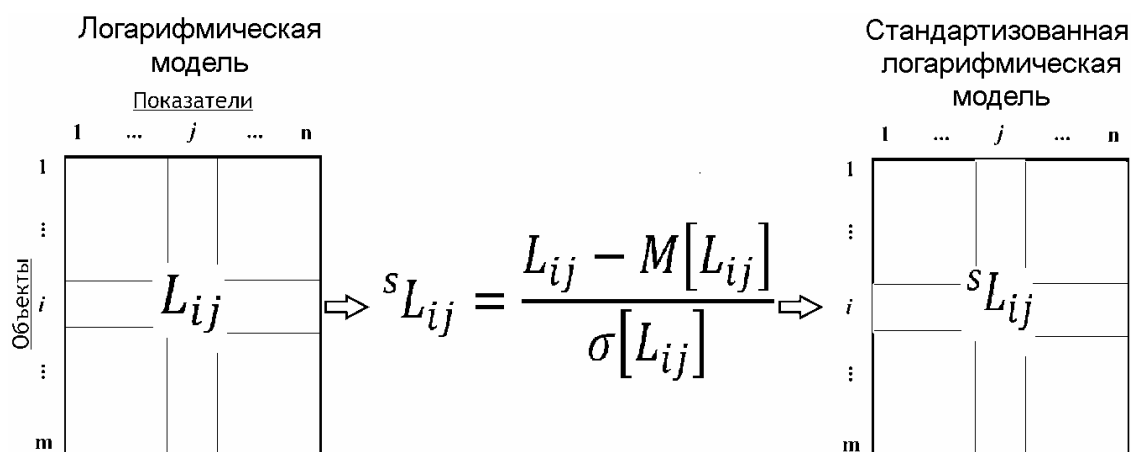


Рис. 4. Схема процедуры стандартизации логарифмической моделей

рованной моделями могут оказаться в различных масштабах (тонна, килограммы, единица условного топлива; теплотворность 1 тонны угля, измеренная в мегаваттах или джоулях). Чтобы обойти это ограничение и перейти к качественному анализу, используется процедура стандартизации (Рис. 4), т.е. приведение частных моделей к единой метрике долей собственных среднеквадратичных отклонений.

Последовательность аналитического процесса на каждом этапе (вход-выход) генерирует матрицу (в некотором случае это может быть вектор-строка или вектор-столбец), где переход к последующему шагу будет зависеть от размерности и типа инверсии матрицы «выхода». Поэтому в семантику записи элементов преобразования необходимо внести дескриптивный атрибут, описывающий «направление» расчета входной матрицы (по столбцам, по строкам или по всей матрице).

Для упорядоченных систем (сигналов) в ЭМА предложена кумулятивная энтропийная модель, схема построения которой представлена на Рис.5. Аналогичным образом вводится кумулятивная логарифмическая и инвертированная модели.

Наряду с тэг-элементами, описывающими семантику математических моделей энтропийного метода, в языке метаописаний EAML были введены две вспомогательные служебные процедуры матричных преобразований: объединения и извлечение столбцов и строк различных матриц, взятых из любых предыдущих шагов в последовательности преобразований, применяемые для подготовки расчетной матрицы к шагу с использованием комбинированных моделей ЭМА. По аналогии с процедурой стандартизации в данных элементах присутствует атрибут «направления» извлечения и объединения по строкам и столбцам (Рис. 6).

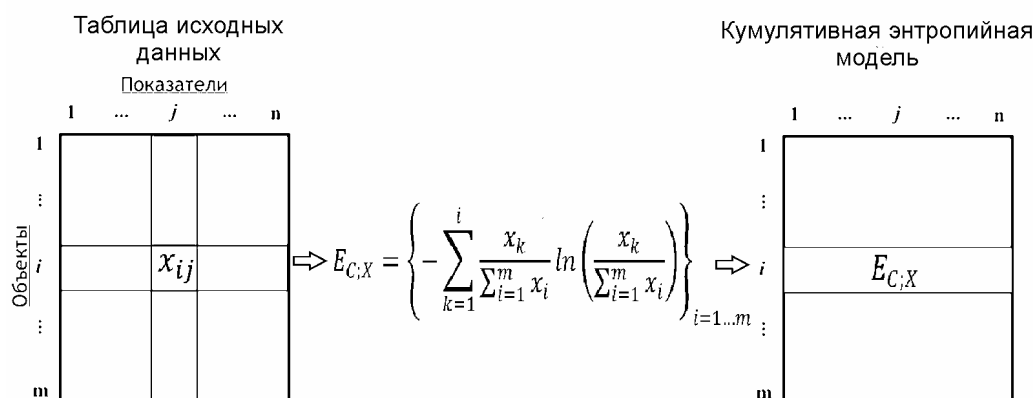


Рис. 5. Схема процедуры аккумуляции энтропийной модели, где $E_{C,X}$ - кумулятивная i -ая вектор строка энтропийной модели

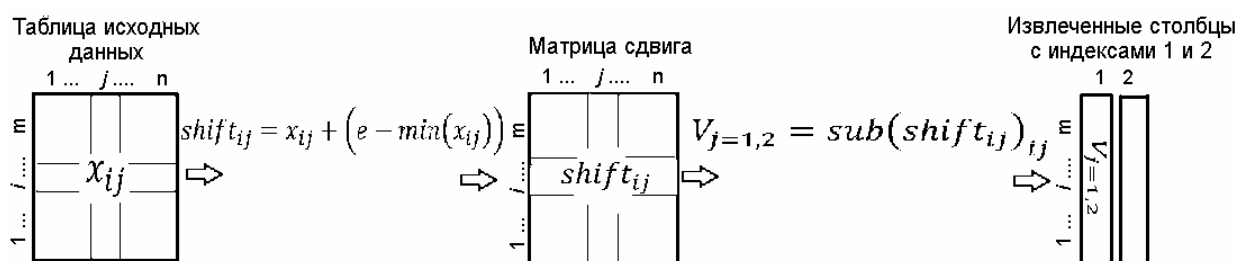


Рис. 6. Пример схемы применения процедур извлечения

Анализ систем с применением ЭМА возможно записать в виде последовательности процедур, использующих определенные выше тэг-элементы. Следовательно, для корректной записи и определения порядка следования тэг-элементов необходимо ввести систему нумерации (идентификаторов), назовем ее "система ссылочных множеств". Суть ее заключается в следующем: каждый элемент должен иметь атрибут равный числовому идентификатору, уникальному для каждого тэг-элемента в последовательности, а также, атрибут-ссылку на числовой идентификатор предыдущих тэг-элементов в последовательности. Причем, один элемент может ссылаться на несколько предыдущих, в этом случае в значение атрибута-ссылки идентификаторы записываются через запятую. Например, как видно из Рис. 6, процедуре объединения потребуется выходной результат матриц энтропийных моделей E_V и E_W , которые являются разными «ветками» расчета.

Главное преимущество ссылочных множеств заключается в возможности ухода от традиционной древовидной модели записи алгоритмов, что облегчает, во-первых, воспри-

ятие системно-аналитического профиля, а, во-вторых, делает процедуру его отладки более прозрачной. Поиск ошибок сводится к последовательному исключению процедур до искомого момента коллизии.

Для элементов, описывающих процедуру извлечения и комбинирования, каждый из которых оперирует с элементами входной матрицы, необходимо указывать какие номера строк или столбцов вовлекать в процедуру комбинирования или извлечения. Для этого был введен дополнительный числовой атрибут, указывающий на номер элемента входной матрицы, атрибутом же «направления» будет определяться столбец это или строка.

Комбинированная модель энтропийного метода анализа разрешает производить операции сложения и/или вычитания, как произвольное количество столбцов, так и строк матрицы, комбинируя различные варианты математических операций сложения или вычитания. То есть, фактически определяется "маска знаков" для преобразования матрицы, где знак (плюс или минус) определяет поэлементную математическую операцию.

Основное семантическое правило составления последовательностей преобразования ЭМА на языке EAML, можно сформулировать следующим образом: порядок и семантика тэг-элементов в последовательности преобразований исходной матрицы полностью подчиняется требованиям теоретического обоснования ЭМА.

Определение фазового пространства предполагает получение для выборочных данных группы связанных границ между видами состояния на основе инвариантных критериев, вытекающих из фундаментальных свойств динамических систем. Поэтому в соответствии с основами теории колебаний, границами состояния на фазовых портретах принимаются плоские кривые, являющиеся изображениями решений дифференциальных уравнений динамических систем с требуемыми (известными) свойствами. В ЭМА используют следующие модели границ (границ видов состояния системы): линейная, эллиптическая и гиперболическая. Совокупность этих моделей определяет фазовую плоскость, где отображаются объекты, с координатами заданными как вектор-столбцы, получаемые в результате вышеописанных математических преобразований. По заданным координатам объектов строятся модели границ.

Поэтому для перехода к анализу на фазовой плоскости в семантику языка EAML были введены элементы, обозначающие области анализа показателей в расчете моделей ЭМА для координат отображаемых объектов в системе объекты-показатели, что дало возможность аналитику оперировать с графическими фазовыми портретами.

Обозначим расчетные области математических моделей ЭМА, как потоки, тогда семантические тэг-элементы примут вид <streamX> и <streamY>. Каждый поток может содержать тэг-элементы математических преобразований, в последовательности согласно основному семантическому правилу языка EAML. Объединяя расчетные потоки в группу, можно ввести понятие расчетной схемы анализа. Схема понимается как совокупность дескрипторов последовательности математических преобразований, заданных тэг-элементами, определяющая фазовые координаты в системе объекты-параметры для отображения конкретных границ видов состояния. Определим соответствующий тэг-элемент как <schema>.

Предполагаемый подход в закладываемой концепции языка EAML заключается в упорядочении и структурировании в единый системно-аналитический профиль (САП) расчетных схем., где профиль понимается, как уникальная сущность, отражающая всю необходимую информацию: матрица "объекты-показатели", реляция к базе данных, если это необходимо, последовательность элементарных математических преобразований для конкретного решающего правила в совокупности действий экспертной системы. Поэтому, целесообразно составлять САП, как свод общих синтаксических правил на расширяемом языке разметки XML. В бета-версии языка EAML авторами предлагается схема DTD, показанная на Рис. 7.

Ввиду ограниченности объема статьи опустим подробное описание XML-элементов САП. Частично расчетные схемы показаны в примере использования EAML в задаче анализа социально-экономических систем.

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE profile [
<!ELEMENT profile {matrix, schema+}>
<!ATTLIST profile
  id ID #REQUIRED
  name CDATA #REQUIRED>
<!ELEMENT matrix {column+}>
<!ATTLIST matrix
  connectionstring CDATA #REQUIRED
  refreshinterval CDATA #REQUIRED
  remotetable CDATA #REQUIRED
  startdatabound CDATA #REQUIRED
  step CDATA #REQUIRED>
<!ELEMENT submatrix {column+}>
<!ATTLIST submatrix
  startdatabound CDATA #REQUIRED
  step CDATA #REQUIRED>
<!ELEMENT column EMPTY>
<!ATTLIST column
  id CDATA #REQUIRED
  name CDATA #IMPLIED
  isobject CDATA #IMPLIED>
<!ELEMENT schema {streamX, streamY, graphexpert?}>
<!ATTLIST schema
  id ID #REQUIRED
  name CDATA #REQUIRED
  desc CDATA #IMPLIED
  refreshinterval CDATA #REQUIRED
  namepattern CDATA #REQUIRED
  legend CDATA #REQUIRED>
<!ELEMENT streamX {submatrix, module*}>
<!ATTLIST streamX
  isnull CDATA #REQUIRED
  axisname CDATA #IMPLIED>
<!ELEMENT module EMPTY>
<!ATTLIST module
  id CDATA #REQUIRED
  mask CDATA #IMPLIED
  direction CDATA #IMPLIED
  parameterindex CDATA #IMPLIED
  parentid CDATA #IMPLIED
  step CDATA #IMPLIED
  name CDATA #REQUIRED>
```

Рис. 7. DTD-модель системно-аналитического профиля (фрагмент)

Регионы РФ (Объекты исследования)	Численность населения (Population)	Численность в городского населения (Pop-city)	Численность сельского населения (Pop- country)	Численность мужского населения (pop-men)	Численность женского населения (Pop-women)	Численность молодежи (Pop-youngs)	Численность трудоспособного населения (Pop- works)	Числен- ность нетрудос- пособного (Pop- pensioners)	Количество преступле- ний против личности (Crime-with- victims)	Местный бюджет - доходы, млн. руб. (Budget- prod)	Местный бюджет - расходы, млн. руб. (Budget- crd)	Среднеду- шевые денежны е доходы (в месяц), руб. (Average- prop-cash)	Среднеме- сячная номиналь- ная начислен- ная заработ- ная плата, руб. (Average- prop-cash)	
Челябинская область	3508	2854	654	1610674	1898059	564293	2193841	750599	19294	582945	52976	53822	14161,2	14829,2
Республика Бурятия	961	527	434	453843	506899	197004	611400	152338	6649	109554	16432	16662	11298,5	14417
Республика Тыва	314	161	153	148432	165508	90644	193649	29647	2611	19776	7492	7578	7871,2	13614,6
Республика Хакасия	538	382	156	250280	287774	96314	343128	98612	2100	64029	8782	9119	10763,9	14488,4
Алтайский край	2497	1341	1156	1155398	1341378	396848	1572851	527077	10447	223751	32502	33747	9748,6	9731,5
Забайкальский край	1117	711	406	533777	583253	228575	712211	176244	5746	113230	21981	21683	10971,5	15142,5
Красноярский край	2890	2187	703	1346790	1542995	485996	1872172	531617	13608	734414	94636	98225	15604,5	18934,7
Иркутская область	2505	1976	529	1159775	1345802	465093	1586464	454020	22783	403031	42410	43767	12881,6	17072,1
Кемеровская область	2822	2397	425	1300622	1521237	462913	1784521	574425	17580	444352	71854	73022	14439,3	15410
Новосибирская область	2640													
Омская область		1992	648	1214454	1425403	405042	1671520	563295	13564	382186	52795	55435	12838,1	15713,6
	2014	1397	617	928300	1085835	325597	1289921	398617	8014	301803	24951	25749	13626,5	13524,8
Томская область	1038	717	321	485179	553329	164841	684738	188929	5490	216059	18953	19972	13481,7	17675,3

Рис. 8. Социально-экономические показатели Сибирского федерального округа [7]

3. Пример использования EAML в задачах анализа данных

Одним из весьма продуктивных направлений развития и применения EAML и его прикладной реализации в РПК «ЭнтропияПлюс» [1] является анализ социально-экономических систем.

Так в одной из работ было проведено исследование состояния регионов Сибирского федерального округа (Челябинская область включена для фона) по набору показателей (Рис. 8), в котором ищались качественные отличия регионов в связи бюджетов, доходов и состава населения.

Результатом первой расчетной схемы (Рис. 9) является гистограмма средних по системе информативностей показателей, представленная в стандартизированном виде. Здесь видно, что экстремальное количество информации имеют: заработная плата – как ведущий информативный показатель; численность сельского населения – стратифицированный показателем с явными лидерами и аутсайдерами. На этом фоне подавленными, но различимыми оказываются аналогичные свойства численности пенсионеров и валового регионального продукта. Также существенной независимостью от остальных показателей отличается и уровень преступности.

Так на фазовом портрете Красноярского края относительно системы выделяются среднедушевой доход и средняя заработная плата – это самый обеспеченный регион (Рис. 10). А высокий уровень линейной связи фазовых координат свидетельствует о подобию структуры показателей в регионе среднему по СФО.

Радикально от системы Красноярского края отличается Алтайский край с высокой негативной связью фазовых координат (Рис. 11). Относительно других показателей он имеет крайне низкую среднюю заработную плату, и заниженный относительно бюджетных показателей среднедушевой доход. Выделяется регион почти аномально (на границе эллипса) высокой численностью сельского населения.

Как «средний» регион выглядит Кемеровская область (Рис. 12). Относительно системы она автономна, так как связь фазовых координат стремится к нулю. Вызвано это сильным недобором средней заработной платы и сельского населения относительно других показателей.

На портретах всех регионов была отмечена композиция (синхронное поведение) численности пенсионеров и валового национального

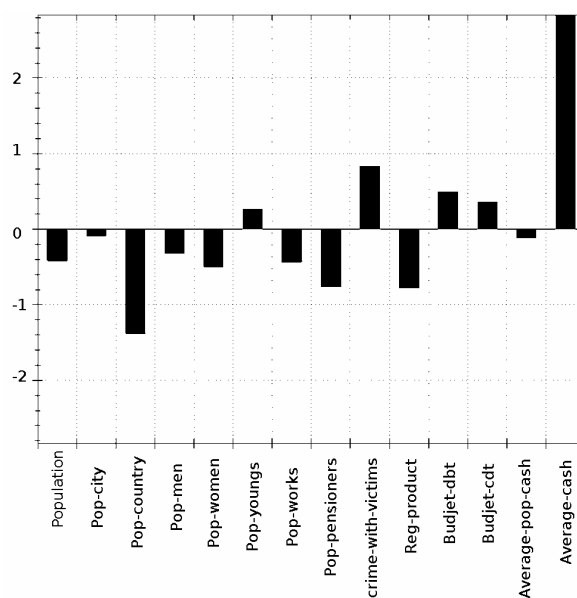


Рис. 9. Гистограмма средних по системе информативностей показателей, представленная в стандартизированном виде

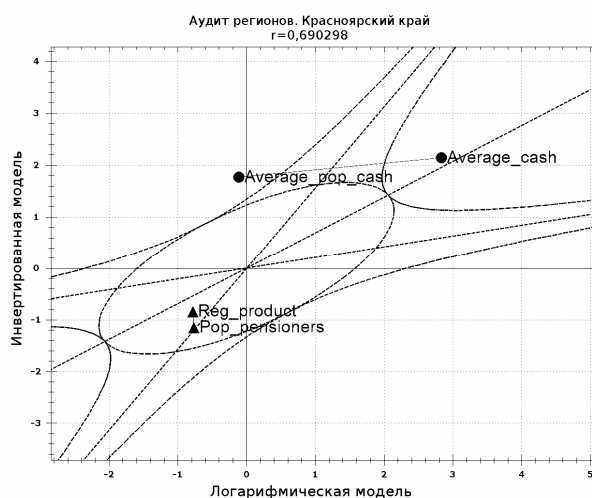


Рис. 10. Фазовый портрет свойств социально-экономических показателей Красноярского края (высокий уровень линейной связи показателей, средняя заработная плата и среднедушевой доход)

продукта (объекты-точки в виде треугольника). Эта особенность подтверждена и в среднем по системе. Трудно представить себе социально-экономические условия и связи, приводящие к такой зависимости. Более вероятной причиной видится некая схема составления отчетности, в которой численность пенсионеров определена как частное пенсионного фонда региона и средней пенсии, а размер пенсионного фонда – как доля регионального продукта.

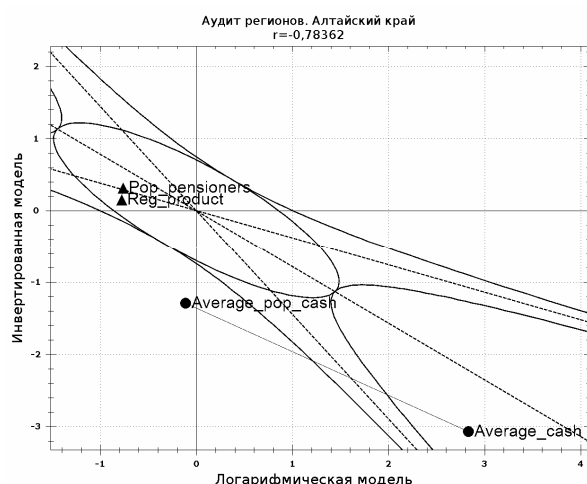


Рис. 11. Фазовый портрет с высокой негативной связью фазовых координат показателей (средняя заработная плата, среднедушевой доход)

Поскольку фазовое пространство энтропийной модели обладает метрикой, то все специфические соотношения и положения относительно границ изображающих точек показателей легко формализуются с помощью аппарата аналитической геометрии. Отсюда имеем возможность составления решающих правил аналитической (экспертной) системы, которые, будучи составленными и проверенными на одном наборе данных, легко переносятся на другие аналогичные по составу таблицы.

Заключение

Результатом работы является оригинальная технология разработки систем интеллектуального анализа данных, основанная на метанаборах описаний модулей решающих и поведенческих правил, и ее прикладная реализация в виде языка EAML.

Заложенный в EAML подход на основе процедурной последовательности позволяет моделировать на уровне понятных терминов (в контексте работы, тэгов) не только решения энтропийного анализа, но и большинства других статистических методов.

Новый подход не только не умаляет достоинств известных языков искусственного интеллекта, но и гарантирует интеграцию различных вычислительных и экспертных модулей на программном уровне с языками C, CLIPS и т.п. Также он позволяет расширить энтропийный

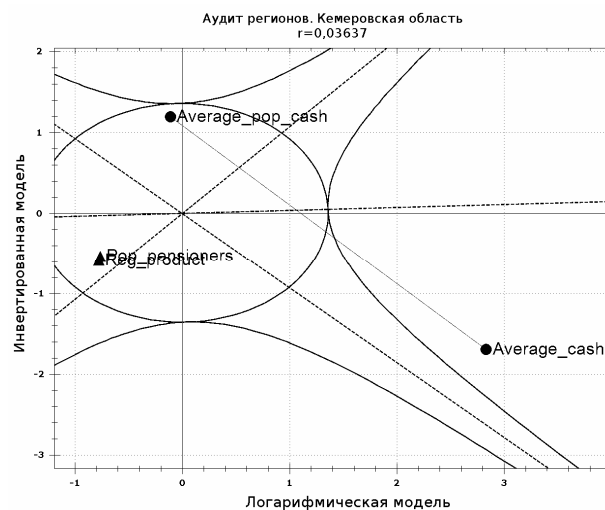


Рис. 12. Фазовый портрет свойств социально-экономических показателей Кемеровской области (связь фазовых координат стремится к нулю, $r \approx 0,03$)

метод в области применения других методов анализа (прикладной статистики, кластерного, корреляционного, регрессионного), что, по мнению авторов, в совокупности может привести к созданию технологии конкурентной Data mining.

Результаты работы предполагается использовать для создания междисциплинарного веб-сервиса по обработке и анализу данных в прикладных и научных исследованиях.

Литература

1. Michalsk, R. S., Kaufman K. Building Knowledge Scouts Using KGL Metalanguage Fundamenta Informaticae, vol. 40, 2000. pp 433-447.
2. Imielinski T., Virmani A., Abdulghani A. DataMine: Application Programming Interface and Query Language for Database Mining, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996 pp.256-261..
3. Частиков А.П., Гаврилова Т.А., Белов Д.Л. Разработка экспертных систем. Среда CLIPS. Санкт-Петербург. БХВ-Петербург. 2003.
4. Замараев Р.Ю., Попов С.Е. Программный комплекс для интеллектуального анализа данных в горном производстве // Информационные технологии и вычислительные системы. 2010. № 3. С. 86-101.
5. Логов А.Б., Замараев Р.Ю., Логов А.А. Моделирование тенденций поведения элементов систем уникальных объектов / Вычислительные технологии, том 10, №5, 2005, – С.54-56.
6. Логов А.Б., Замараев Р.Ю., Логов А.А. Анализ функционального состояния промышленных объектов в фазовом пространстве. Институт угля и углехимии СО РАН, Кемерово: 2004 .168 с.

7. Сайт Федеральной службы государственной статистики // [Электронный ресурс]. URL: <http://www.gks.ru/dbscripts/munst/munst.htm> (дата обращения 10.02.2011).

Попов Семен Евгеньевич. Старший научный сотрудник Института вычислительных технологий Сибирского отделения РАН. Окончил Кемеровский государственный университет в 2002 году. Кандидат технических наук. Автор 20 печатных работ. Область научных интересов: математическое моделирование систем и процессов, распределенные системы и GRID-технологии. E-mail: ogidog@mail.ru

Замараев Роман Юрьевич. Старший научный сотрудник Института вычислительных технологий Сибирского отделения РАН. Окончил Кузбасский государственный технический университет в 1996 году. Кандидат технических наук. Автор 37 печатных работ, в том числе четырех монографий. Область научных интересов: математическое моделирование систем и процессов, анализ данных. E-mail: zrukem@gmail.com