Информационные процессы, Том 11, № 3, 2011, стр. 378–393

ИНФОРМАЦИОННОЕ ВЗАИМОДЕЙСТВИЕ

Мониторинг многокомпонентных систем: предметно-независимые модели и методы

В.П.Супин

Институт проблем передачи информации им. А.А. Харкевича, Российская академия наук, Москва, Россия

Поступила в редколлегию 25.09.2011

Аннотация—В работе рассматриваются модели и методы мониторинга многокомпонентных систем, обобщающие практический опыт построения информационных систем мониторинга. Для представления прикладной области использована модель асинхронной многокомпонентной системы, зависимость состояния которой от времени описана набором гипотез. Гипотезы проверяются на основании сигналов от датчиков, считывающих состояние отдельных компонент. Достоверность выбора той или иной гипотезы определяется достоверностью информации, которою несут сигналы, поступающие от датчиков. Предложены методы, позволяют идентифицировать повторяющиеся и противоречивые сигналы, а также сгруппировать сигналы, несущие информацию о независимых информационных процессах, проистекающих одновременно. Особенностью предложенного подхода является выделение формальной модели знаний предметной области и методов, оперирующих абстрактными математическими объектами, что позволяет применять их к задачам мониторинга в различных предметных областях.

1. ВЕДЕНИЕ

В широком смысле, под мониторингом понимают идентификацию отклонений характеристик объекта мониторинга от штатного режима функционирования, либо выявления тенденции изменения характеристик на основе систематического сбора и анализа информации об объекте. Задача мониторинга характерна для тех видов деятельности, где изучаемый объект представляет сложную многокомпонентную систему, о состоянии которой можно судить по изменению свойств системы в целом или совокупности составляющих её объектов. В качестве примеров видов деятельности, неотъемлемой частью которых является задача мониторинга, можно привести эпидемиологический надзор, управление крупномасштабными техническими объектами и производствами, общественная безопасность.

Сбор данных является неотъемлемой частью мониторинга, поэтому задачам сбора данных, посвящено большое количество работ, рассматривающих как технологические, так и теоретические (например, [1]–[4]) основы решения данной задачи. В простейших случаях, для которых достаточно решить задачу сбора данных, сигнал датчика является одновременно и сигналом об отклонении режима функционирования системы от штатного режима. Пожарная сигнализация является очевидным примером, когда срабатывание одного датчика является сигналом об отклонении условий на наблюдаемом объекте от штатного режима. Возможностей систем сбора данных недостаточно для мониторинга систем, штатный режим которых характеризуется наличием “фоновых” сигналов датчиков. Наличие фона характерно для большинства сложных систем в экономике, медицине, технике. Наглядным примером служит мониторинг заболеваемости, где сигналом о выходе из штатного режима является превышение порога заболеваемости.

Для мониторинга таких систем необходимы более точные методы обработки информации, работающие в условиях наличия фона. Достаточной для применения в широком диапазоне предметных областей общностью обладают методы интеллектуального анализа данных, исследуемые в рамках искусственного интеллекта. Однако их применение на практике часто ограничено в связи со сложностью реализации и обучения интеллектуальных структур. Пробел между простейшими методами и методами интеллектуального анализа данных приводит к появлению большого количества узкоспециализированных прикладных систем, число которых даже в одной предметной области достаточно велико. В частности, по оценкам, приведённым в работе Д. Бравата [5], только в области эпидемиологии насчитывается не менее 115 систем мониторинга, по которым опубликовано более семнадцати тысяч научных работ.

В данной работе предложены методы обработки информации, использующие подходы интеллектуального анализа данных для обработки сигналов с датчиков системы мониторинга.

2. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ МНОГОКОМПОНЕНТНОЙ СИСТЕМЫ

В общем виде исследуемая система состоит из множества объектов, взаимодействующих между собой, и множества датчиков, измеряющих параметры объектов и передающие их в центр обработки данных (рис. 1).

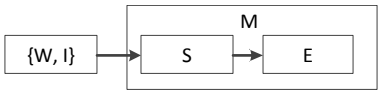

Рис. 1. Общая схема исследуемой системы.

На схеме введены следующие обозначения: $\{W, \mathrm{I}\}$ - многокомпонентная система, система мониторинга $\mathrm{M} = \{\mathrm{S}, \mathrm{E}\}$. Многокомпонентная система $\{W, \mathrm{I}\}$ представлена множеством $\mathrm{W} = \left\{w_1, \ldots, w_{|\mathrm{W}|}\right\}$ объектов мониторинга и моделью изменения состояния объектов I. Система мониторинга $\mathrm{M} = \{\mathrm{S}, \mathrm{E}\}$ состоит из множества датчиков системы $\mathrm{S} = \left\{\mathrm{s}_1, \ldots, \mathrm{s}_{|\mathrm{S}|}\right\}$ и центра обработки данных E. Прямые скобки в нижних индексах обозначают размерность соответствующих множеств. Объект $w \in W$ описывается вектором свойств $\mathbf{D} = \left(d_1, \ldots, d_{|\mathbf{D}|}\right)$, его размерность $|\mathbf{D}|$ зависит от решаемой задачи. Элементы $d_i \in R$, $i = 1 \ldots D$ вектора $\mathbf{D}$ описывают внутреннее состояние объекта.

Характерной чертой современных подходов к моделированию сложных систем является формулирование гипотез, описывающих динамику процессов на микроуровне, ведущих к пониманию динамики основных агрегированных характеристик на макроуровне [6]. Модель I описывает динамику изменения состояния $\mathbf{D}$ отдельных объектов $w \in W$ во времени и может быть представлена в следующем виде. Состояние объекта W описывается вектором $\mathbf{D}$, и изменяется в некоторые дискретные моменты в соответствии с правилом:

$$d_{i,\,new} = a_{i1}d_1 + a_{i2}d_2 + \ldots + a_{i|\mathbf{D}|}d_{|\mathbf{D}|} + f_i,$$

где a_i – коэффициенты вектора соответствующих размерностей, а f_i – элемент вектора $\mathbf{f}$ внешнего воздействия на объект W. Обозначим через $\ldots < T_0 < T_1 < \ldots < T_n < \ldots$ моменты изменения состояния объекта W. Тогда изменение состояния $d_i(t)$ объекта W может быть описано уравнением

где предполагается постоянство функции $d_i(t)$ на каждом интервале $T_n < t \leq T_{n+1}$. Уравнение описывает изменение состояние объекта, обусловленное внутренними законами его функционирования и внешним воздействием $\mathbf{f}$. При наличии взаимодействия между объектами

$w \in W$ уравнение для состояния $\mathbf{D}_i$ объекта w_i принимает следующий вид

$$\mathbf{D}_i\left(T_{n+1}\right) = \mathbf{a}_{ii}\mathbf{D}_i\left(T_n\right) + \sum_{j=1,\, i\neq j}^{|W|} \mathbf{a}_{ij}\mathbf{D}_j\left(T_n\right) + \mathbf{f}_i(T_n),$$

где $\mathbf{a}_{ii}\mathbf{D}_i\left(T_n\right)$ характеризует внутренние законы изменения состояния объекта, $\sum\limits_{j=1,\, i\neq j}^{|W|} \mathbf{a}_{ij}\mathbf{D}_j\left(T_n\right)$ описывает взаимодействие объектов, а $\mathbf{f}_i(T_n)$ описывает воздействие внешней среды на объект w_i. В ряде случаев может быть применена запись соответствующих уравнений в дискретном виде [7]. Подытоживая сказанное, можно сказать, что модель изменения состояний объектов I состоит из следующих элементов:

$$I = \{A,\, F\},$$

где A – матрица, составленная соответствующим образом из векторов $\mathbf{a}_{ij}$, F – матрица внешних воздействий на объекты w системы.

В общем случае, моменты изменения состояний различных объектов T не совпадают, то есть система является асинхронной. Уравнения, записанные в матричной форме, могут быть исследованы методами анализа асинхронных систем (см. например [8] – [10]). Для исследования могут применяться вероятностные и другие методы. С точки зрения мониторинга систем и процессов интересны агрегированные характеристики системы, отражающие динамику процессов на макроуровне.

Агрегированные показатели, такие как уровень заболеваемости, преступности и т.д., часто отражают зависимость величин подмножеств множества объектов W от времени. Подмножества вводятся по принципу близости вектора состояния $\mathbf{D}$ объекта к центральному вектору подмножества:

$$w \in C \Longleftrightarrow \rho_C\left(\mathbf{D},\, \mathbf{D}_C\right) < \varepsilon_C,$$

где w – объект из множества объектов W, $C \subset W$ – подмножество множества объектов W, ρ_C – расстояние между векторами состояний объектов, $\mathbf{D}$ – вектор состояния объекта w, $\mathbf{D}_C$ – центральный вектор подмножества C, ε_C – радиус подмножества C. Таким образом, подмножество C задаётся следующим набором

$$C = \{\rho_C,\, \mathbf{D}_C,\, \varepsilon_C\}.$$

Каждому объекту w может соответствовать одно или более подмножеств, в этом случае будем говорить, что объект принадлежит одному или более классам:

$$C\left(w\right) = C(\mathbf{D}) = \{C_{w1},\, \ldots\},$$

где $\mathbf{D}$ – вектор состояния объекта w поскольку $\mathbf{D} = \mathbf{D}(t)$, то и набор классов объекта меняется во времени, то есть $C\left(w\right) = C\left(t\right)$. Тогда величина $|C|$ класса C определяется как количество объектов w, принадлежащих классу C: $|C| = \sum\limits_{w\in W} w \in C$, где выражение $w \in C$ принимает значение 1, если объект w принадлежит классу C и 0 в противном случае. В терминах объектов и классов агрегированные показатели, характеризующие динамику процессов на макроуровне, принимают вид

$$F\left(t\right) = F\left(\left|C_1(t)\right|,\, \left|C_2(t)\right|,\, \ldots,\, \left|C_n(t)\right|,\, t\right),$$

где $F\left(t\right)$ – агрегированный показатель состояния системы, $C_1,\, \ldots,\, C_n$ – классы объектов, t – время. В данном выражении функциональная зависимость F от величин классов $|C|$ обозначает как зависимость от самой функции $|C\left(t\right)|$, так и от её производных, интегральных и других возможных характеристик.

Сформулируем задачи мониторинга состояния системы Wв терминах предложенной математической модели. Будем рассматривать две задачи: задачу идентификации отклонения характеристик системы от штатного режима и задачу проверки гипотез о характере отклонения.

Задача идентификации отклонения характеристик системы от штатного режима заключается в выборе таких набора классов $C_1, \ldots, C_n$, функции $F(t)$ методов её измерения и штатного режима $F_0(t)$и порога отклонения ε_F, что отклонение $|F(t) - F_0(t)| > \varepsilon_F$ будет соответствовать существенной коррекции модели изменения состояний объектов I. Понятие существенного отклонения зависит от конкретной области применения модели и может быть формализована, например, в терминах устойчивости асинхронной системы W.

Задача проверки гипотез о характере отклонения характеристик системы от штатного режима расширяет задачу идентификации отклонения характеристик системы. Помимо идентификации отклонения $|F(t) - F_0(t)| > \varepsilon_H$ необходимо выбрать наиболее вероятную гипотезу $H_{prob}(t)$ из множества гипотез $H = \{H_1, \ldots, H_N\}$ такую, что

$$|F(t) - H_{prob}(t)| = \min_{H_i \in H}(|F(t) - H_i(t)|).$$

Выигрыш от проверки гипотез развития процесса заключается в выборе значения $\varepsilon_H < \varepsilon_F$ и, как следствие, раннем обнаружении отклонения характеристик от штатного режима при доле ложных срабатываний меньшей, чем можно получить простым уменьшением порога ε_F в выражении $|F(t) - F_0(t)| > \varepsilon_F$. На практике в качестве нормы $|\bullet|$, применяемой для обнаружения отклонения характеристик систем, целесообразно использовать интегральную характеристику $|\int_{\Delta t}(F(t) - F_0(t))|$ по интервалу времени Δt для подавления всплесков $F(t)$.

3. ОПИСАНИЕ СУЩЕСТВУЮЩИХ ПРОБЛЕМ, СИСТЕМ, В КОТОРЫХ ОНИ НАБЛЮДАЮТСЯ, И ПУТИ ИХ РЕШЕНИЯ

В формулировках задач мониторинга, приведённых выше, определены абстрактные элементы модели, которые необходимо связать с концепциями предметной области при проектировании той или иной системы мониторинга: множество объектов W, вектор состояния объекта $\mathbf{D}$, классы объектов $C_1, \ldots, C_n$, функцию $F(t)$ агрегированных характеристик системы W, функцию $F_0(t)$ штатного режима функционирования системы W, норму $|\bullet|$ и порог ε_F отклонения $F(t)$ от штатной функции $F_0(t)$, а также альтернативные гипотезы $H = \{H_1, \ldots, H_N\}$ и пороговое значение ε_H. К задачам, решаемым при построении системы мониторинга, также относятся задача планирования измерений (например, [11]) и задача выбора алгоритма опроса датчиков (например, [1]–[4]).

Задача выбора нормы $|\bullet|$ отклонения $F(t)$ от штатной функции $F_0(t)$ рассматривается в целом ряде работ ([12], [13] и др.). Из рассматриваемых в работах методов можно отметить те, которые обладают достаточной степенью общности для применения в различных предметных областях: метод Сёрфлинга (Serfling method); авторегрессионное интегрированное скользящее среднее (Autoregressive Integrated Moving Average, ARIMA); рекурсивный метод наименьших квадратов (Recursive Least Square, RLS); экспоненциально взвешенное скользящее среднее (Exponentially Weighted Moving Average, EWMA); метод накопленных сумм (Cumulative Sums, CUSUM); скрытые Марковские модели (Hidden Markov Models, HMM); вейвлет-анализ (Wavelet algorithms); обобщённая линейная смешанная модель (Generalized Linear Mixed Modeling, GLMM); методы на основе Байесовских сетей (Bayesian networks); метод опорных векторов (Prospective Support Vector Clustering, PSVC).

Отметим, что точность перечисленных методов непосредственно зависит от качества данных о состоянии системы, что обуславливает существование задачи обработки собранных дан-

ных о системе. Существует два подхода к получению информации о состоянии системы объектов W. Один заключается в получении информации о состоянии всех объектов системы одновременно и часто применяется при ситуационном мониторинге, то есть при необходимости единовременно получить "срез" состояния системы W. Разновидностью методов ситуационного мониторинга являются процедуры на основе репрезентативных выборок, нашедшие широкое применение в социальных науках. Основным свойством репрезентативной выборки, позволяющим распространить результаты измерения на генеральную совокупность, заключается в отражении значимых свойств генеральной совокупности. Методам построения репрезентативных выборок посвящено значительное количество исследований ([17]). Другой метод к получению информации о состоянии системы W применяется при непрерывном мониторинге. Он заключается в инкрементном накоплении информации об объектах $w \in W$ по мере их взаимодействия с датчиками системы. Этот метод широко применяется при исследовании интересов аудитории интернет-ресурсов и построении профиля пользователя интернет-ресурса. Проникновение в повседневную жизнь технологий электронной коммуникации позволяет использовать метод и для мониторинга состояний систем, далёких от интернета. Примером может служить исследование тенденций распространения гриппа Google ([18]).

Измерение состояния $\mathbf{D}(t)$ объекта w системы производиться при помощи датчиков S, результат работы которых $\hat{\mathbf{D}}(t)$ является сигналом об изменении состояния объекта:

$$E(t_j) = \begin{cases} 1 : \rho\left(\hat{\mathbf{D}}(t_j) - \hat{\mathbf{D}}(t_{j-1})\right) > \varepsilon_D \\ 0 : \text{else} \end{cases},$$

где функция $E(t_j)$ принимает значение 1 в те моменты времени t_j, когда обнаружено значительное изменение свойств объекта; t_j – моменты времени, в которые производились измерения состояния объекта; ε_D – порог, характеризующий значимость изменения. Информацию о событии E несёт сигнал $\hat{\mathbf{D}}(t)$, производимый датчиком, поэтому наблюдаемое время t_j возникновения события зависит от принципа работы датчиков и их набора. Изменение состояния $\hat{\mathbf{D}}$ объекта w может быть обнаружено несколькими датчиками в разные моменты времени. Каждый датчик сгенерирует сигнал о событии, в результате чего получим последовательность сигналов об одном и том же событии:

$$E(t_j) = 1, E(t_{j+1}) = 1, \ldots, E(t_{j+n}) = 1,$$

где n – количество датчиков, обнаруживших изменение состояния $\hat{\mathbf{D}}(t)$ объекта. Существует несколько причин, которые определяют "размножение" события, среди них:

- Состояние $\hat{\mathbf{D}}$ объекта w измеряют несколько датчиков: S_1, S_2, S_3, ...;
- Датчики S_i и S_j измеряют разные характеристики объекта w, то есть получают разный набор входных данных: $\mathbf{D} = \{\mathbf{D}_i, \mathbf{D}_{ij}, \mathbf{D}_j\}$;
- Датчики i и j производят разный набор результирующих значений: $\hat{\mathbf{D}} = \left\{\hat{\mathbf{D}}_i, \hat{\mathbf{D}}_{ij}, \hat{\mathbf{D}}_j\right\}$;
- События E_i и E_j похожим образом изменяют вектор состояния объекта $\hat{\mathbf{D}}$.

В результате, вместо одного момента времени t для каждого события получаем n значений, что искажает наблюдаемую функцию $F(t)$ и, как следствие, приводит к сбоям при обнаружении отклонения состояния системы от штатного режима и при прогнозировании поведения системы.

Решение проблемы искажения функции $F(t)$, возникающего вследствие "размножения" сигналов о событии, связано с решением двух задач:

1. Идентификации повторных и противоречивых сигналов: Для сигналов о событиях $E(t_i)$ и $E(t_j)$ установить, сигнализируют они об одном и том же событии, или о разных;
2. Кластеризации сигналов по их принадлежности к различным протекающим в системе информационным процессам.

Проблема характерна для систем мониторинга, в которых состояние многокомпонентной системы измеряется множеством взаимосвязанных датчиков. На рис. 2 представлена UML-диаграмма ([16]) объектной модели системы мониторинга, в которой наблюдается обозначенная проблема.

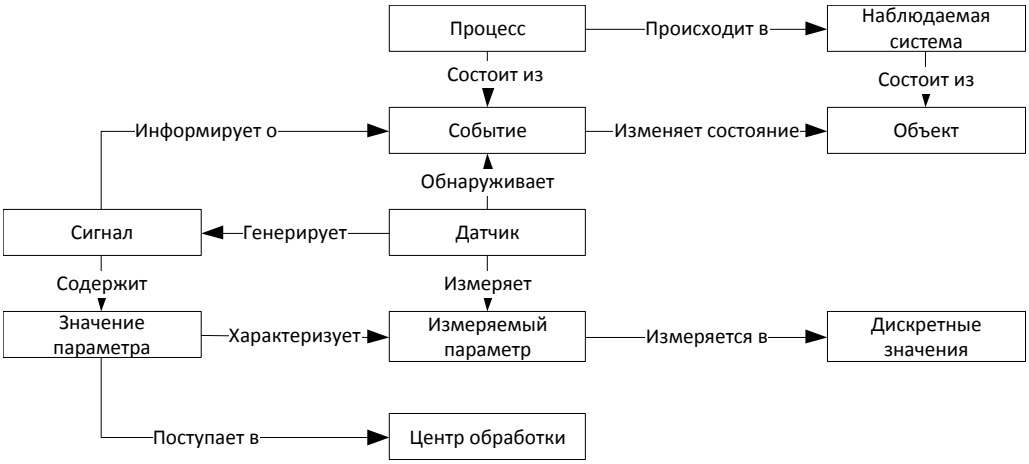

Рис. 2. UML-диаграмма объектной модели системы мониторинга входами (справа).

Соответствие между UML-диаграммой и математической моделью задаётся в таблице 1. Объектная модель используется при проектировании программного обеспечения прикладных систем мониторинга. В последней главе представлены примеры систем мониторинга, для каждой элементы модели описаны в терминах предметной области.

Таблица 1. Соответствие элементов UML-диаграммы и переменных математической модели входами (справа)

Наблюдаемая система	W
Объект	w
Процесс	$F(t)$
Событие	E
Датчик	S
Измеряемые параметры	$\mathbf{D}$
Множество значений измеряемых параметров	
Сигнал	$E(t)$
Множество значений выходных параметров	$\hat{\mathbf{D}}$

В следующей главе для решения обеих задач, связанных с размножением сигналов, будут предложены соответствующие процедуры. Обе реализованы с использованием алгоритма выбора расширенного пространства значений скрытой части состояния объекта.

4. ОПИСАНИЕ МЕТОДА ИДЕНТИФИКАЦИИ ПОВТОРЯЮЩИХСЯ И ПРОТИВОРЕЧИВЫХ СИГНАЛОВ

Для обнаружения и исправления систематических ошибок измерения состояния $F(t)$ системы W требуется привлечение знаний предметной области. Это может быть достигнуто как разработкой специальных алгоритмов для узких предметных областей, так и использованием обучаемых интеллектуальных структур. В первом случае тиражируемость решения ограничена, так как алгоритмы включают специфику предметной области и наработки передаются только с опытом разработчика. При использовании обучаемых интеллектуальных структур требуется значительное количество обучающих примеров. Их разработка должна выполняться совместно специалистами предметной области и специалистами в области искусственного интеллекта. В результате обученная модель также применяется только к решению узкой задачи. Вербализация ([14], [15]) модели, которая бы позволила её модифицировать, требует дополнительного исследования.

Проблемой описанных методов является то, что формализованные знания предметной области заложены в алгоритмы обработки информации, что делает невозможным их простую адаптацию к новым задачам. Ниже предлагаются алгоритмы обработки данных, которые взаимодействует с формализованной моделью знаний предметной области, для решения задач, определяемых данной моделью (рис. 3). На рисунке введены следующие обозначения: A – ал-

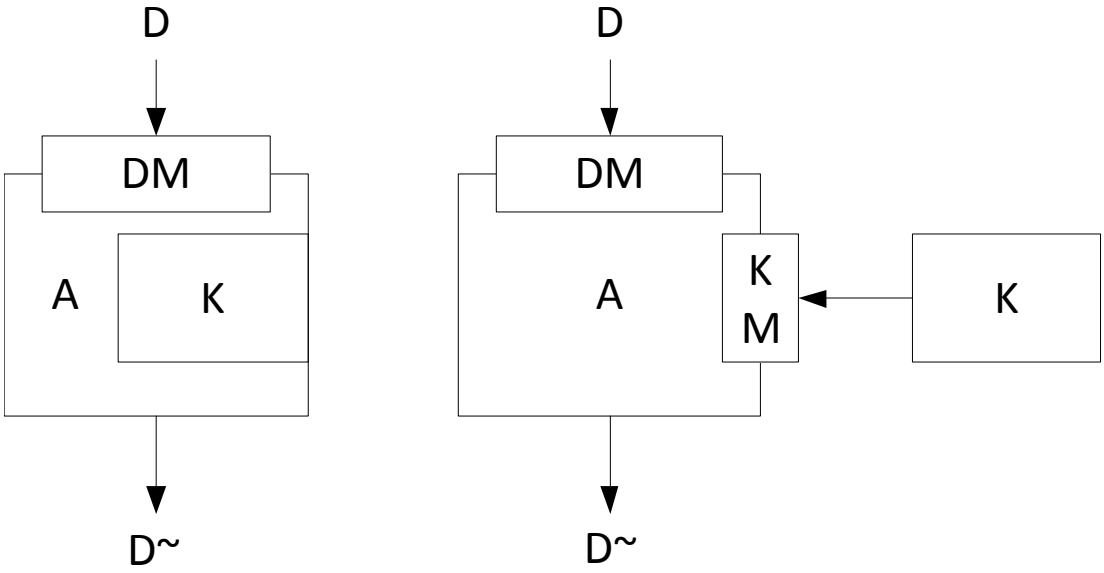

Рис. 3. Традиционные алгоритмы (слева) и новые алгоритмы с двумя входами (справа).

горитм обработки данных; D – входные данные алгоритма; D~ - выходные данные алгоритма; DM – интерфейс входных данных (модель данных); K – знания предметной области; KM – интерфейс модели знаний. Принципиальным отличием предлагаемой структуры алгоритмов интеллектуального анализа и обработки информации является выделение модели знаний предметной области в независимый блок и взаимодействие алгоритма с этим блоком средствами формализованного интерфейса модели знаний.

На рис. 4 представлена общая схема системы мониторинга, пунктирной линией выделена часть, в которой предметно-независимым способом решаются задачи идентификации повторных и противоречивых сигналов и кластеризации сигналов по их принадлежности к информационным процессам. Указанные задачи решаются в блоках “Преобразование сигналов в

события” и “Объединение событий в процессы”. Блоки “Обнаружение отклонения от штатного режима” и “Построение прогноза развития”, получают на вход данные, обработанные процедурами устранения ошибок.

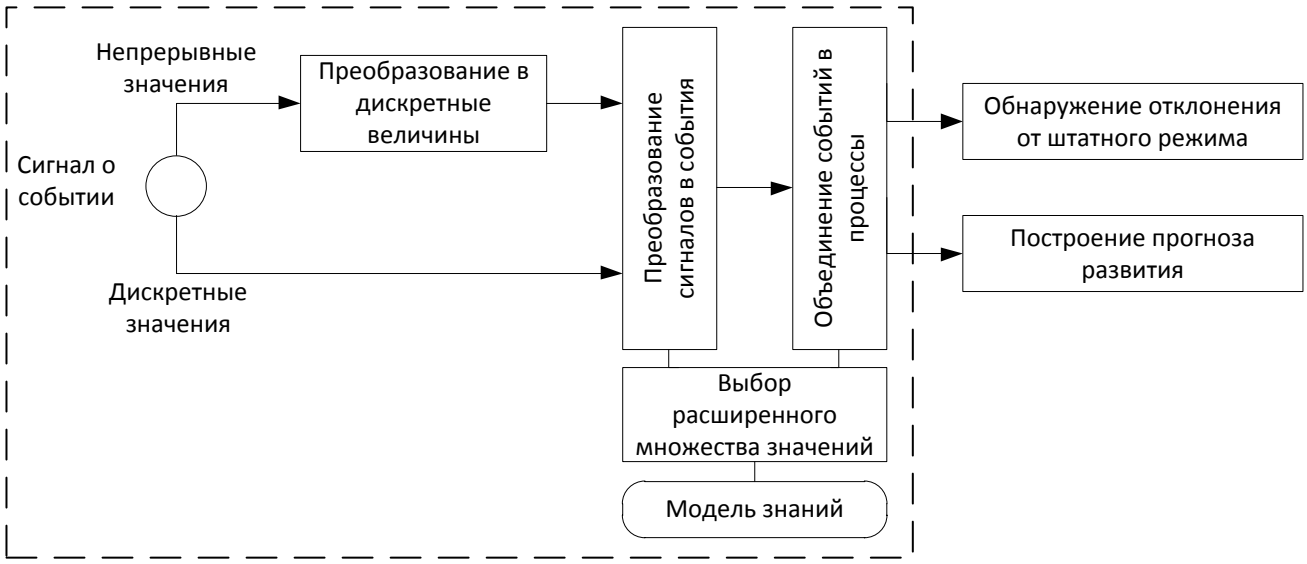

Рис. 4. Общая схема решения входами (справа).

Ниже приводится описание процедур предметно-независимого блока. Описаны процедуры, обозначенные на схеме, а также алгоритмы и модели, необходимые для этих процедур:

1. Преобразование непрерывных входных параметров в дискретные значения;
2. Мера близости векторов состояний объектов;
3. Выбор формальной модели представления знаний предметной области;
4. Алгоритм выбора расширенного множества значений состояния объекта;
5. Преобразование сигналов в события;
6. Объединение событий в процессы;

Преобразование непрерывных входных параметров в дискретные значения. Данные о состоянии объекта $\mathbf{D}$ должны быть представлены в формализованном виде. Для этого должны быть выбраны шкалы и единицы измерения. В случае, когда элемент d_i вектора $\mathbf{D}$ является численным результатом измерения, то единицы измерения и шкалы известны. Если для описания значений элемента d_i используются лингвистические конструкции, то требуется их формализация. Действительно, для описания некоторых явлений, которые не принято или невозможно измерять численными значениями, используются такие термины, как термины “наблюдается”, “незначительный”, “не наблюдается”.

Наблюдаемые параметры могут быть описаны непрерывными или дискретными значениями. И вектор значений $\mathbf{D}$ состоит из дискретной и непрерывной составляющих:

$$\mathbf{D}=\left\{\mathbf{D}^{\text{discr.}}, \mathbf{D}^{\text{cont.}}\right\}.$$

Дискретным значениям ставятся в соответствие узлы семантической сети. Для этого задаётся отображение множества значений каждой из составляющих $d_i \in \mathbf{D}^{\text{discr.}}$ Дискретной составляющей вектора $\mathbf{D}$ на узлы семантической сети V:

$$D_i \to V,$$

где D_i – счётное и конечное множество значений i-ой компоненты d_i вектора $\mathbf{D}^{\text{discr.}}$. В общем случае отображение может не быть взаимно однозначным. Например, нескольким элементам множества D_i может соответствовать один элемент множества V. Этот случай имеет место, когда используется недостаточно подробная семантическая сеть, либо значения множества D_i содержит "синонимы": $\exists d_j,\ d_k \in D_i : d_j \approx d_k$.

Отображение непрерывных значений $\mathbf{D}^{\text{cont.}}$ на узлы семантической сети требует дополнительного шага, так как множество V узлов сети является дискретным множеством. Непрерывные значения должны быть поделены на диапазоны, каждый из которых соответствует узлу семантической сети. Для представления диапазонов используются нечёткие множества, при этом выбор характеристической функции $\mu(d_i^{cont.})$ зависит от решаемой задачи. В случае, когда нечёткие множества имеют пересечения, происходит активизация нескольких узлов семантической сети.

Мера близости векторов состояний объектов. Входные данные, представленные дискретными значениями, либо преобразованные к дискретным значениям способом, описанным выше, становятся входными данными процедуры "Преобразование сигналов в события". Процедура решает задачу идентификации и устранения повторяющихся и противоречивых сигналов о состоянии объекта многокомпонентной системы, для чего использует меру близости сигналов о событиях E_i и E_j. Сигнал о событии содержит информацию $\hat{\mathbf{D}}(t)$ о состоянии объекта в момент времени t, поэтому, мера близости сигналов зависит от меры близости векторов состояний объектов. Введём меру близости $\rho(\mathbf{D}_i, \mathbf{D}_j)$ векторов состояний объектов $\mathbf{D}_i(t)$ и $\mathbf{D}_j(t)$ и выберем алгоритм её измерения.

Состояние объекта $\hat{\mathbf{D}}$ описывается наблюдаемой $\hat{\mathbf{D}}^{obs}$ и скрытой $\hat{\mathbf{D}}^{hdn}$ частями, между которыми существует вероятностная связь:

$$\hat{\mathbf{D}}=\left\{\hat{\mathbf{D}}^{\text{obs}};\hat{\mathbf{D}}^{\text{hdn}}\right\},\ P(\hat{\mathbf{D}}^{\text{hdn}})=P(\hat{\mathbf{D}}^{\text{hdn}}|\hat{\mathbf{D}}^{\text{obs}}),$$

где $P\left(\hat{\mathbf{D}}^{\text{hdn}}\right)$ – вероятность того, что скрытое состояние $\hat{\mathbf{D}}^{\text{hdn}}$ соответствует действительному состоянию объекта $\mathbf{D}^{\text{hdn}}$. При этом считаем, что наблюдаемое состояние объекта $\mathbf{D}^{obs}$ измерено без ошибок, т.е.

$$P(\hat{\mathbf{D}})=P(\hat{\mathbf{D}}^{\text{hdn}}|\hat{\mathbf{D}}^{\text{obs}})P(\hat{\mathbf{D}}^{\text{obs}})=P(\hat{\mathbf{D}}^{\text{hdn}}|\hat{\mathbf{D}}^{\text{obs}}).$$

Таким образом, задача сводится к измерению близости векторов состояний $\hat{\mathbf{D}}^{hdn}$ при условии получения результатов измерения $\hat{\mathbf{D}}^{obs}$. Введём меру близости векторов $\hat{\mathbf{D}}_i^{hdn}$ и $\hat{\mathbf{D}}_j^{hdn}$ в следующем виде:

$$\rho(\hat{\mathbf{D}}_i^{\text{hdn}},\hat{\mathbf{D}}_j^{\text{hdn}})=\min\left[P(\hat{\mathbf{D}}_i^{\text{hdn}}\mid\hat{\mathbf{D}}^{\text{obs}});P(\hat{\mathbf{D}}_j^{\text{hdn}}\mid\hat{\mathbf{D}}^{\text{obs}})\right],$$

следовательно,

$$\rho(\mathbf{D}_i,\mathbf{D}_j)=\rho(\mathbf{D}_i^{\text{hdn}},\mathbf{D}_j^{\text{hdn}})=\min\left[P(\mathbf{D}_i^{\text{hdn}}\mid\mathbf{D}^{\text{obs}});P(\mathbf{D}_j^{\text{hdn}}\mid\mathbf{D}^{\text{obs}})\right].$$

На практике данная мера близости будет мала для тех пар векторов состояний $\bar{S}$, скрытые части $\bar{S}^{hdn}$ которых с близкой вероятностью проявляются при наблюдаемых измеренных значениях $\bar{S}^{obs}$.

Выбор формальной модели представления знаний предметной области. Условная вероятность $P(\mathbf{D}^{hdn}|\mathbf{D}^{obs})$ зависит от семантики элементов векторов $\mathbf{D}^{obs}$ и $\mathbf{D}^{hdn}$ и описывается информационной моделью знаний предметной области. Следовательно, формальная модель знаний предметной области должна иметь средства представления условной вероятности –

причинно-следственных связей. Это свойство присутствует у большинства существующих моделей знаний: формальной логики, нейросетей, семантических графов и пр. Для описания информационной модели выбрано формальное представление в виде графа семантической сети по ряду причин:

1. Для представления вероятностных характеристик причинно-следственных связей в системах искусственного интеллекта часто применяют Байесовскую сеть доверия, в основе которой лежит ациклический направленный граф;
2. Байесовская сеть доверия, как и другие структуры искусственного интеллекта, требуют процедуры обучения. Семантическая сеть позволяет задать структуру графа, сведя обучение к выбору условных вероятностей;
3. Для целого ряда областей знаний уже существуют онтологии в виде семантических сетей, которую можно использовать в виде исходных данных для обучения

Граф семантической сети состоит из вершин, называемых концептами, и рёбер, представляющих отношения между концептами:

$$G = (C, R),$$

где C – множество концептов, R – множество отношений между ними. Для использования семантической сети для определения меры близости векторов $\mathbf{D}^{hdn}$, вводится отображение между элементами пространства значений векторов $\mathbf{D}^{hdn}$ и множеством узлов C графа семантической сети G:

$$\mathbf{D}^{\mathrm{hdn}} \to C^{\mathrm{hdn}}, C^{\mathrm{hdn}} \subset C,$$

аналогичное отображение применяется и к элементам вектора $\mathbf{D}^{obs} \to C^{obs}$. Задача обучения Байесовской сети состоит в выборе структуры связей между элементами множеств C^{obs} и C^{hdn}. Под структурой связей понимается набор вспомогательных узлов C графа G, выборе связей между ними и присвоении каждой связи значения условной вероятности.

Алгоритм выбора расширенного множества значений состояния объекта. Использование графа семантической сети в качестве основы построения Байесовской сети позволяет свести процедуру обучения к выбору значений условных вероятностей для рёбер графа сети, имея в наличии набор узлов и связей между ними. Таким образом, задача обучения семантической сети разделяется на две подзадачи:

1. Первоначальное обучение, включающее выбор онтологии предметной области и её подстройку при помощи экспертов;
2. Корректировка вероятностных коэффициентов Байесовской сети на основе данных, появляющихся при использовании созданной с использованием системы мониторинга.

Первоначальное обучение семантической сети начинается с выбора графа G, представляющего онтологию предметной области из существующих, либо разработке необходимой онтологии и представлении её в виде графа. Необходимым условием применения графа G, представляющего онтологию предметной области, для оценки меры близости векторов является возможность сопоставления $\mathbf{D}^{hdn} \to C^{hdn}$ и $\mathbf{D}^{obs} \to C^{obs}$. То есть вершины графа G должны содержать множество, соответствующее наблюдаемым и скрытым параметрам и их дискретным значениям.

Граф семантической сети должен выражать причинно-следственные связи между концептами, поэтому необходимым шагом подготовки графа онтологии является устранение циклов. Отражение причинно-следственных связей предполагает, что граф G является направленным

графом, поэтому результатом устранения циклов станет граф, являющийся суперпозицией древовидных графов.

$$G = T_1 \cup T_2 \cup \ldots.$$

При использовании в качестве основы для построения графа G онтологии предметной области, отдельные деревья T_i в составе графа G представляют ни что иное, как таксономии, элементами которых являются элементы множеств C^{hdn}, C^{obs}, а также элементы, соответствующие промежуточным уровням классификации (рис. 5).

На рисунке заштрихованными кругами показаны узлы графа, соответствующие наблюдаемым параметрам – "причинам" в терминах причинно-следственной связи. Сплошными кругами показаны скрытые состояния – "следствия". Изогнутые линии – рёбра графа, которым в результате обучения должна быть назначена условная вероятность. Рёбра графа, изображённые прямыми и пунктирными линиями, соответствуют различным деревьям, входящим в граф.

Назначение условных вероятностей рёбрам графа происходит следующим образом. Предположим, что $c \in C^{hdn}$, а $c_i \in C^{obs}$, $i = 1 \ldots n$ – узлы, соответствующие "причинам" узла c.

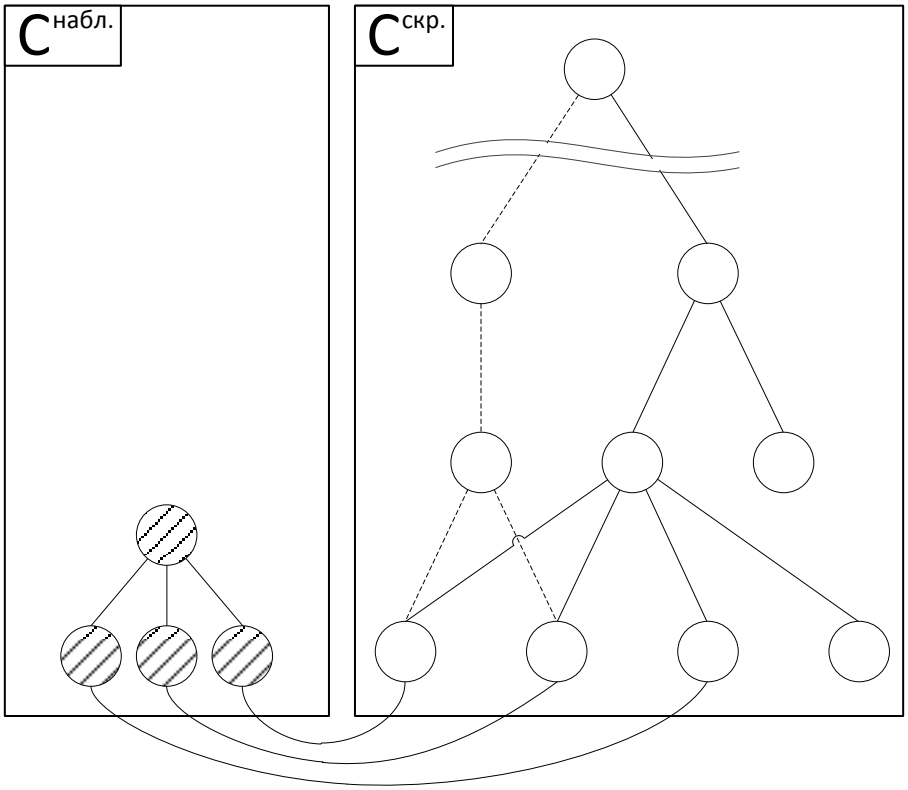

Рис. 5. Граф семантической сети, представленный в виде суперпозиции древовидных структур.

В случае если наблюдаемые параметры c_i попарно независимы, то вероятность события c определяется по формуле полной вероятности:

$$p\,(c) = \sum_i \mathbf{D}^{\mathrm{obs}}[i] p(c|c_i),$$

где $\mathbf{D}^{obs}[i]$ равняется 1, если i-ый параметр наблюдается и 0 иначе. В случае, когда между наблюдаемыми параметрами c_i существует зависимость, обучение потребует записи условной вероятности $p(c|\bigcup c_i)$, где $\bigcup c_i$ – всевозможные комбинации событий c_i.

При использовании данного подхода требуется значительное количество примеров для обучения сети, которое определяется числом перестановок элементов c_i для каждого элемента $c \in C^{hdn}$. Сложность возрастает, если принять во внимание тот факт, что состояние некоторых параметров $c_i \in C^{obs}$ может приобретать не только два значения – 1 или 0, но и значение "неизвестно".

С другой стороны, граф G представлен в виде суперпозиции древовидных графов T_i, при этом T_i являются таксономиями, то есть классификациями концептов сходными признаками. Это обозначает, что чем меньше расстояние между узлами дерева T, тем ближе расположены соответствующие векторы состояний $\mathbf{D}$:

$$\rho(\mathbf{D}_i^{\text{hdn}}, \mathbf{D}_j^{\text{hdn}}) \sim l(C_i^{\text{hdn}}, C_j^{\text{hdn}}),$$

где $l(C_i^{hdn}, C_j^{hdn})$ – расстояние между узлами графа C_i^{hdn}, C_j^{hdn}, соответствующим скрытым состояниям $\mathbf{D}_i^{hdn}$, $\mathbf{D}_j^{hdn}$.

В дальнейшем, мера близости $\rho(\mathbf{D}_i^{hdn}, \mathbf{D}_j^{hdn})$ будет использоваться для кластеризации состояний $\mathbf{D}^{hdn}$, то есть необходимо выбрать параметр ε такой, что при

$$\rho(\mathbf{D}_i^{\text{hdn}}, \mathbf{D}_j^{\text{hdn}}) < \varepsilon$$

векторы $\mathbf{D}_i^{hdn}$, $\mathbf{D}_j^{hdn}$ попадают в один кластер. При этом деревья T в составе графа G, основанного на онтологии, представляют таксономию – классификацию концептов. То есть, чем ближе находится общий родитель двух концептов, тем более близки векторы $\mathbf{D}^{hdn}$, которые они представляют.

Таким образом, для первоначального обучения семантической сети при помощи эксперта меру близости векторов $\rho(\mathbf{D}_i^{hdn}, \mathbf{D}_j^{hdn})$ можно записать в виде:

$$\rho(\mathbf{D}_i^{\text{hdn}}, \mathbf{D}_j^{\text{hdn}}) = l(C_i^{\text{hdn}}, C_j^{\text{hdn}}) = \min\left(l(C_i^{\text{hdn}}, C_p), l(C_j^{\text{hdn}}, C_p)\right),$$

где C_p – общий "родитель" узлов C_i^{hdn} и C_j^{hdn}. Если общего родителя нет, то $\rho(\mathbf{D}_i^{hdn}, \mathbf{D}_j^{hdn}) = \infty$. При определении общего родителя C_i^{hdn} и C_j^{hdn} поиск производится только на множестве

$$E(C_i^{\text{hdn}}, C_j^{\text{hdn}}) = E(C_i^{\text{hdn}}) \cap E(C_j^{\text{hdn}}),$$

где $E(C^{hdn})$ – расширенное множество значений элемента C^{hdn}, определяемое как множество наследников узла, находящегося на n узлов выше C^{hdn}:

$$E(C^{\text{hdn}}) = succ(par(C^{\text{hdn}}, n)),$$

где $par(C, n)$ – множество всех родителей узла C, находящихся n узлами выше. Множество может иметь размерность больше 1, так как узел $C \in G$ может входить в несколько деревьев T; $succ(C)$ – множество всех "наследников" узла C во всех деревьях T в которые он входит.

Число n для каждого узла C определяется для каждого вхождения узла C в дерево T на основании экспертной оценки, либо в зависимости от связей узла C^{hdn} с узлами C^{obs}:

$$P(\bar{C}^{hdn}|\bar{C}^{obs}).$$

Экспертная оценка заключается в выборе эмпирической зависимости числа n от значений вектора состояния $\mathbf{D}^{obs}$

$$n = n(\mathbf{D}^{obs}).$$

Использование экспертной оценки числа n позволяет использовать онтологию предметной области в качестве основы для Байесовской сети в случае, когда для оценки условных вероятностей $P(\mathbf{D}^{hdn}|\mathbf{D}^{obs})$ не существует достаточно данных для обучения сети.

Преобразование сигналов в события. Расстояние $\rho(\mathbf{D}_i^{hdn}, \mathbf{D}_j^{hdn})$ между векторами состояний $\mathbf{D}^{hdn}$, введённое указанным выше способом, позволяет установить меру близости состояний $\mathbf{D}_i$, $\mathbf{D}_j$ одного или более объектов w. Сигнал о событии несёт также информацию о времени события и данные $\mathbf{I}$, идентифицирующие источник события - объект w:

$$\{\mathbf{I}, \mathbf{D}, t\}.$$

В случае, когда сигналы $\mathbf{D}_i$, $\mathbf{D}_j$ несут информацию о состоянии одного и того же объекта w, то есть

$$\rho(\mathbf{I}_i, \mathbf{I}_j) < \theta,$$

возникает задача идентификации и устранения повторяющихся и противоречивых сигналов о состоянии объекта многокомпонентной системы. В противном случае, решается задача кластеризации данных для идентификации независимых процессов, протекающих в распределённой многокомпонентной системе.

Процедура идентификации и устранения повторяющихся и противоречивых сигналов позволяет исключить влияние сигналов, поступающих повторно, от различных датчиков или несущих недостоверную информацию о событии.

При поступлении сигнала, вектору скрытого состояния $\mathbf{D}^{hdn}$ ставится в соответствие расширенное множество значений E, величина которого в соответствии с алгоритмом выбора расширенного множества значений тем больше, чем меньше вероятность $P(\mathbf{D}^{hdn}|\mathbf{D}^{obs})$:

$$|E| \sim n \sim \frac{1}{P(\mathbf{D}^{\mathrm{hdn}}|\mathbf{D}^{\mathrm{obs}})}.$$

Величина расширенного множества значений $|E|$ характеризуют дисперсию измеренного состояния объекта $\mathbf{D}^{hdn}$. Сигналы, для которых $|E|$ превышает заданное пороговое значение, считаются недостоверными.

При поступлении двух (или более) сигналов $\mathbf{D}_i$, $\mathbf{D}_j$, расширенные множества значений которых пересекаются, принимается решение об идентичности сигналов. Все идентичные сигналы, кроме одного, устраняются как дубли. Выбор одного из множества идентичных сигналов определяется либо минимальной дисперсией, либо свойствами, зависящими от условий задачи – самый новый, и т.д. Сигнал, представляющий множество идентичных сигналов, а также его дисперсию, будем называть событием.

Объединение событий в процессы. События подвергаются процедуре объединения в процессы. События представлены сигналами, для которых $\rho(\mathbf{I}_i, \mathbf{I}_j) > \theta$, то есть сигналы несут информацию о разных объектах w_i и w_j. Процессом называется множество событий, связанных общими свойствами. Для объединения сигналов в процессы также используется алгоритм выбора расширенного множества значений с незначительными дополнениями.

Решение об объединении событий в процесс принимается при выполнении следующих условий:

$$\begin{cases} \rho(\mathbf{D}_i, \mathbf{D}_j) > \theta \\ E(\mathbf{D}_i^{\mathrm{hdn}}, \mathbf{D}_j^{\mathrm{hdn}}) \neq \emptyset \\ \rho(t_i, t_j) < \tau \end{cases}.$$

Последнее условие отражает близость событий по времени, при этом τ – характерное время события, зависящее от вида события:

$$\tau = \tau(\mathbf{D}^{hdn}).$$

При объединении событий в процессы можно использовать меру близости $\rho_{\{T\}}(\mathbf{D}_i^{hdn}, \mathbf{D}_j^{hdn})$, которая отличается от описанной выше меры $\rho(\mathbf{D}_i^{hdn}, \mathbf{D}_j^{hdn})$ тем, что вычисление происходит только по подмножеству $\{T\}$ деревьев, составляющих граф G.

5. ПРАКТИЧЕСКОЕ ПРИЛОЖЕНИЕ РЕЗУЛЬТАТОВ

Предложенные процедуры и алгоритмы были использованы при проектировании и разработке систем мониторинга для эпидемиологии и транспортной безопасности. В таблицах приведено соответствие между объектами модели и понятиями предметной области, а также сформулированы задачи идентификации повторных и противоречивых сигналов и кластеризации сигналов по их принадлежности к информационным процессам в терминах предметной области. С использованием предложенных методик разработаны программные комплексы мониторинга.

5.1. Пример 1: Эпидемиология

Соответствие элементов модели и объектов предметной области для примера 1 приведено в таблице 2. Задача идентификации повторных и противоречивых сигналов, а также задача кластеризации сигналов по их принадлежности к протекающим в системе информационным процессам формулируются следующим образом:

Задача 1: Для экстренных сообщений E_i, E_j установить, сообщают они об одном и том же случае заражения человека, либо о разных. Практическая задача состоит в том, что в разных экстренных извещениях одно заболевание может быть описано разными диагнозами.

Задача 2: Для экстренного извещения E установить, к эпидемическому процессу какого заболевания $F(t)$ относиться событие, описываемое извещением.

Таблица 2. Соответствие элементов модели и объектов предметной области для примера 1

Наблюдаемая система	Население города, района, области или страны
Объект	Человек, проживающий на наблюдаемой территории
Процесс	Эпидемиологическая обстановка в стране, представленная количеством больных по каждому из инфекционных диагнозов, а также динамика изменения числа больных
Событие	Заражение человека инфекционным заболеванием.
Датчик	ЛПУ / Территориальное отделение ЦГИЭ
Измеряемые параметры	Измеряемые параметры представлены количественными измерениями – анализами и качественными измерениями – наблюдениями врачей
Множество значений измеряемых параметров	
Сигнал	Экстренное извещение, посылаемое врачом, обнаружившим заболевание, в территориальное отделение ЦГИЭ
Множество значений выходных параметров	Элементы классификатора заболеваний МКБ-10

5.2. Пример 2: Ространснадзор

Соответствие элементов модели и объектов предметной области для примера 1 приведено в таблице 3. Задача идентификации повторных и противоречивых сигналов, а также задача кластеризации сигналов по их принадлежности к протекающим в системе информационным процессам формулируются следующим образом:

Задача 1: Для протоколов E_i, E_j о проверках на объекте транспорта установить, описывают

они одно и то же нарушение, либо разные. Практическая задача состоит в том, что разные контролирующие органы могут квалифицировать нарушение по разным статьям.
Задача 2: Выявить тенденции роста $F(t)$ числа нарушений Е и, как следствие, определить риски и мероприятия по их снижению.

Таблица 3. Соответствие элементов модели и объектов предметной области для примера 2

Наблюдаемая система	Система транспортных предприятий
Объект	Предприятие, входящее в транспортную систему
Процесс	Состояние транспортной системы, представленное в терминах количества нарушений установленных законом нормативов состояния транспорта и организации переворзок
Событие	Нарушение правил эксплуатации транспорта или правил перевозок
Датчик	Подразделение РТН, ответственное за проведение проверок по отдельному виду транспорта на закреплённой за ним территории
Измеряемые параметры	Измеряемы параметры и их возможные значения регламентируются внутренними документами Ространснадзора
Множество значений измеряемых параметров	
Сигнал	Протокол, содержащий результаты мероприятия по проверке транспортного объекта
Множество значений выходных параметров	Номера статей КоАП и других законодательных актов

СПИСОК ЛИТЕРАТУРЫ

1. Маликова Е.Е., Цитович И.И. Стратегия группового поллинга в широкополосных беспроводных сетях мониторинга. Обозрение прикладной и промышленной математики. 2010. Т.17, № 2. С. 284–285.
2. Маликова Е.Е. Метод повышения пропускной способности систем телеметрии и мониторинга на базе беспроводных сетей. T-Comm — Телекоммуникации и транспорт. 2010. № 7. С. 37–39.
3. Маликова Е.Е. Задача оценки параметров качества сетей сбора и обработки телеметрической информации. Труды 64 научной сессии, посвящённой дню радио. Российское научно–техническое общество радиотехники, электроники и связи имени А.С. Попова. Москва. 2009. С. 363–365.
4. Маликова Е.Е., Цитович И.И. Метод группового поллинга при независимой активности датчиков в сетях мониторинга. Информационные процессы. 2011. Т. 11, № 2. С. 291–303.
5. Dena M. Bravata, Kathryn M. McDonald, Wendy M. Smith, Chara Rydzak, Herbert Szeto, David L. Buckeridge, Corinna Haberland, Douglas K. Owens. Systematic Review: Surveillance Systems for Early Detection of Bioterrorism-Related Diseases. Ann Intern Med. 2004. Vol. 140. Pp. 910-922.
6. А. Г. Владимиров, Н. А. Гречишкина, В. С. Козякин, Н. А. Кузнецов, А. В. Покровский, Д. И. Рачинский. Асинхронность: теория и практика. Информационные процессы. 2011. Т. 11, № 1. С. 1–45
7. Baudet G. M. Asynchronous iterative methods for multiprocessors // J. Assoc. Comput. Mach. 1978. Vol. 25, no. 2. Pp. 226–244.
8. Bertsekas D. P., Tsitsiklis J. N. Parallel and Distributed Computation. Numerical Methods. Englewood Cliffs. NJ: Prentice Hall, 1989. 715 pp.
9. Асарин Е. А., Козякин В. С., Красносельский М. А., Кузнецов Н. А. Анализ устойчивости рассинхронизованных дискретных систем. М.: Наука, 1992. 408 с.
10. Kaszkurewicz E., Bhaya A. Matrix diagonal stability in systems and computation. Boston, MA: Birkhauser Boston Inc., 2000. xiv+267 pp. ISBN: 0-8176-4088-6.

11. Малютов М.Б. Нижние границы для средней длины последовательного планирования экспериментов. Известия вузов. Математика. 1983. Т. 27, № 11. С. 19–41.

12. Moore A., Cooper G., Tsui R., Wagner M.. Summary of Biosurveillance-relevant technologies. 2003.

13. Kaustav D. Detecting anomalous records in categorical datasets. 2007. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.

14. А. Н. Горбань, В. Л. Дунин-Барковский, А. Н. Кирдин и др. Нейроинформатика., В кн.: Нейроинформатика. Новосибирск: Наука. 1998. ISBN 5-02-031410-2

15. Gorban A., Mirkes Eu., Tsaregorodtsev V. Generation of Explicit Knowledge from Empirical Data through Pruning of Trainable Neural Networksю. Washington DC. 1999. Proceedings of IJCNN'99. IEEE. Vol. 6. Pp. 4393-4398.

16. Booch G., Jacobson I., Rumbaugh J. OMG Unified Modeling Language Specification. 2000.

17. Рабочая книга социолога. М.: Наука, 1977. С. 257-296

18. Ginsberg J., Mohebbi1 M., Patel R., Brammer L., Smolinski M., Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009.