СИСТЕМЫ ОБЪЕДИНЕНИЯ ДАННЫХ ИЗ РАЗНЫХ ИСТОЧНИКОВ: ПРИНЦИПЫ РЕАЛИЗАЦИИ И АРХИТЕКТУРА ОБРАБОТКИ ДАННЫХ ДЛЯ ОБУЧЕНИЯ СИСТЕМ ПРИНЯТИЯ РЕШЕНИЙ

В. В. Самойлов

Санкт-Петербургский институт информатики и автоматизации РАН
199178, Санкт-Петербург, 14-я линия В.О., д.39
samovl@iias.spb.su

УДК 681.3
В. В. Самойлов. **Системы объединения данных из разных источников: Принципы реализации и архитектура обработки данных для обучения систем принятия решений.** // Труды СПИИРАН, Вып. 1, т. 2. — СПб: СПИИРАН, 2002.
Аннотация. *Обсуждаются проблемы, возникающие при разработке и реализации систем принятия решения, использующих распределенные источники данных и предлагаются пути их решения. Обсуждаются и сравниваются основные принципы, а также предлагается архитектура совместной интеллектуальной обработки распределенных данных для обучения системы.* — Библ. 6 назв.

UDC 681.3
V. V. Samoilov. **Data Fusion Systems: Principles and Architecture for Data Processing in Decision Making System Learning.** // SPIIRAS Proceedings, Issue 1, v. 2. — SPb: SPIIRAS, 2002.
Abstract. *The paper discusses the design and implementation issues of data fusion decision making systems and proposes the basic principles and architecture for distributed data processing aiming at decision making system learning on the basis of knowledge discovery from data.* — Bibl. 6 items.

Введение

Задачи объединения данных, полученных из различных источников, с целью принятия решений возникли первоначально в области приложений военного характера, однако в настоящее время они находят все большее применение во многих прикладных областях. Под объединением данных обычно понимают “процесс комбинирования данных из различных источников таким путем, что результат обеспечивает пользователя той информацией, которая не может быть получена анализом каждого отдельного источника в отдельности” [2]. Ключевой особенностью этой задачи обработки данных является то, что источники данных являются *физически разделенными* и, возможно, *гетерогенными* по структуре. Рассмотрим более подробно эти особенности.

Под физически разделенными источниками данных понимаются данные, размещенные на носителях, которые физически разделены в пространстве. В приложении к компьютерным системам анализа и объединения данных это означает, что источники данных, как минимум, находятся в различных базах данных, а чаще всего и на разных хостах сети.

Источники данных являются гетерогенными по структуре, если состав и структуры данных различных источников различаются. Различные источники могут содержат данные разного вида, например, статистические данные, географические данные, символьные выражения, представленные на различных языках и т.д.

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Кроме того, совместная обработка таких данных затрудняется тем, что данные, полученные из различных источников, могут обладать различной степенью точности, достоверности, полноты и непротиворечивости [3–5].

Отмеченные особенности делают задачу объединения данных весьма нетривиальной, в частности, требуют принципиально новых решений и подходов к их совместной обработке, в особенности тогда, когда целью объединения данных является процесс принятия решений.

1. Проблемы возникающие при объединении данных

Объединение данных является позволяет использовать больше информации, описывающей какую либо сложную ситуацию, состояние сложного объекта или какую-либо иную сторону конкретной предметной области. Очевидно, что в этой задаче речь идет об объединении данных, относящихся к одной и той же предметной области или к области, полученной в результате интеграции различных областей, имеющей некоторый практический смысл. Задача, которую необходимо решать при объединении данных, должна определяться в терминологии, принятой в предметной области, при этом *единый словарь предметной области единое понимание терминов предметной области* являются главными связующими и объединяющими элементами задачи объединения данных, ключом к единому пониманию компонент задачи, задачи в целом и к пониманию результатов ее решения различными экспертами и конечными пользователями.

Однако на практике различие терминологического словаря описания предметной области и трактовки этих терминов экспертами (пользователями) различных локальных источников данных является первой проблемой возникающей при попытке объединения таких данных в рамках единой задачи. Если эта проблема не решена, то задача может быть либо не решена вообще или, что еще хуже, будет решена неверно. Это может получиться как следствие того, что одна и та же характеристика сущности в различных источниках будет соотноситься с различными понятиями предметной области или наоборот, данные, относящиеся к различным понятиям, будут объединяться в рамках одного понятия предметной области.

Следующей проблемой, возникающей при объединении данных, является так называемая проблема идентификации сущности (*entity identification problem*). Суть этой проблемы заключается в необходимости установления одно-однозначного соответствия между описаниями, находящимися в различных локальных источниках, но относящихся к одному экземпляру сущности предметной области.

Рассмотрим эту, во многих случаях основную, проблему более подробно. Пусть на уровне онтологии предметной области определены какие-то сущности, понятия и отношения между ними. В каждом из локальных источников существует своя структура для хранения определенного подмножества сущностей этой онтологии. Каждый локальный источник хранит в рамках своей структуры сведения о нескольких экземплярах сущностей. Для уникальной идентификации экземпляра сущности в локальном источнике данных используется система первичных ключей. С точки зрения предметной области может существовать сколь угодно много экземпляров сущностей онтологии, причем данные об одном и том же экземпляре сущности (прецеденте) могут одновременно находится в различных локальных

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

источниках данных, имея в них свои различные уникальные ключи. Более того, в различных локальных источниках может различаться и сама структура ключей, идентифицирующих данную сущность. При объединении данных необходимо каким либо образом разрешать подобную неопределенность и тем самым получать возможность сводить вместе данные о конкретном экземпляре сущности (прецеденте) из различных локальных источников.

Еще одной проблемой, с которой приходиться встречаться при попытках объединения данных – это проблема различного атрибутивного состава локальных источников. Эта проблема заключается в том, что одна и та же предметная область в различных источниках может описываться различным набором атрибутов. Таким образом, прямое объединение становится невозможным. Более того, эти данные могут быть различны по своей природе, например в одном источнике собирается информация статистического вида, в другом – данные, которые содержат результаты неточных по природе и сильно зашумленных измерений, в третьем источнике данные могут быть представлены в виде символьных выражений на каком-либо языке представления, в четвертом – знания определенной группы экспертов в виде базы знаний и т.п.

Существует также проблема различного представления атрибутов в локальных источниках. Суть этой проблемы заключается в том, что один и тот же атрибут предметной области в различных источниках данных может быть представлен в шкалах различного типа. Принципиальное отличие от предыдущей проблемы заключается в том, что здесь идет речь об одном и том же атрибуте предметной области, но представленном в разных источниках. Кроме того, атрибут в этих источниках может быть измерен в разных масштабах и с разной точностью.

Все вышеперечисленные проблемы являются составными частями одной из основных проблем возникающей при построении систем принятия решений на основе распределенных источников информации (Data Fusion System), которая называется *проблемой неконгруэнтности данных* [2]. Многие эксперты в области построения таких систем отмечают, что по сложности разрешения эта проблема стоит на первом месте. Конечно, при построении таких систем существуют и другие проблемы, но они менее сложны для разрешения и не рассматриваются в данной работе.

2. Анализ возможных схем объединения данных для обучения

Вопрос о выборе путей и схем объединения данных самым непосредственным образом связан с *проблемой многофункциональной интеграции*. В системах, которые непосредственно обрабатывают данные каждого из отдельных источников, могут использоваться различные алгоритмы. Знания, получаемые на основе этих данных, могут, соответственно, быть выражены в различной форме. Системе объединения таких данных и знаний в целях принятия решений должна с такой задачей справляться.

Рассмотрим формальную постановку задачи объединения данных.

Компоненты задачи и структура их взаимодействия представлена на рис. 1. В предметной области *Data domain* фиксируются некие события *ID1 … IDN*. Каждое такое событие имеет какую-либо *единственную* интерпретацию $CI_1 \ldots CI_N$. Прогнозирование интерпретации вновь наступивших событий и является целью системы объединения данных. Интерпретация может быть

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

выражена как в виде структуры классов, так и в числовых шкалах. В системах принятия решений традиционно оперируют понятием класса ситуации, и именно такая постановка рассматривается в данной работе. Предметная область реально представлена набором распределенных источников данных *DS1...DSk.* В каждом из них фиксируется некоторое подмножество событий ID_{DS}. В общем случае выделяется подмножество событий, зафиксированных в каждом из источников ID_{SH}. В ядре системы существует своя база знаний с набором закономерностей {R}, выявленных при обучении на мета-уровне. Кроме этого, для принятия решения об интерпретации события *ID* в системе принятия решений существует набор *Мета-классификаторов* {*MC*}, [1] т.е — настроенных механизмов принятия решения над базой знаний системы, объединенных единой схемой принятия решений. Таким образом, формально функция системы объединения данных представляется в виде:

$$\left(\{DS\}, ID_{SH}\right) \Rightarrow MC\_CI\,,$$

где DS_n — источник данных *N*; *ID* — уникальный идентификатор события; MC_G_CI — класс состояния события, спрогнозированный *G*-м механизмом принятия решения на мета-уровне.

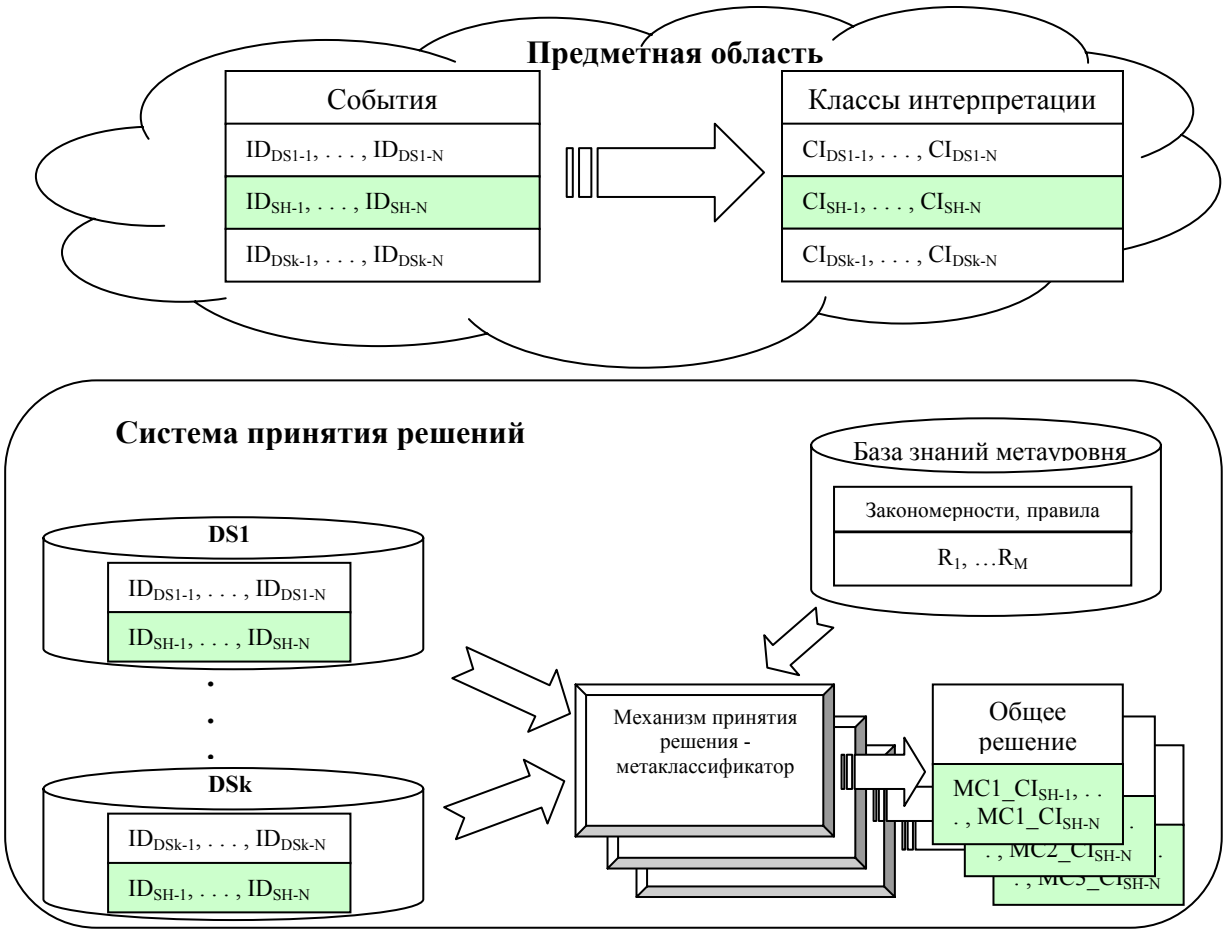

Рисунок 1 Соответствие реальной предметной области и системы принятия решений на основе распределенных источников данных.

Очевидно, что $\{MC\_CI\} \in \{CI\}$ и $MC_G\_CI_{SH\text{-}k}=CI_{SH\text{-}k}$ если для события *SH-k* система с помощью *G*-го механизма принятия решения правильно определила интерпретацию *CI* этого события.

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Структура объектов источника данных более подробно представлена на рис. 2.

В источнике данных *DSk* зафиксирован набор событий {*ID*}. Фиксация событий происходит в виде набора атрибутов наблюдаемых объектов *Сущность.Атрибут* (*Entity.Attribute)*. В каждой локальной системе принятия решений существует своя локальная база знаний с набором выявленных при обучении закономерностей *{R}.* Кроме этого для принятия решения об интерпретации события *ID* в локальной системе принятия решений существует набор *Базовых классификаторов {BC}* – настроенных механизмов принятия решения над локальной базой знаний системы [1]. Формально функция системы принятия решения локального источника представляется в виде

$$\left(\{E.A\}, ID_{DS \cup SH}\right) \Rightarrow BC\_CI\ ,$$

где *E.A* – значение атрибута *A* сущности *E*; *ID* – уникальный идентификатор события; BCg_{DSk}_CI – класс состояния события, спрогнозированный *G*-м механизмом принятия решения источника данных *DSk*. Очевидно также, что, как и в предыдущем случае, $\{BC\_CI\} \in \{CI\}$ но не всегда $BCg_{DSk}\_CI=CI_{DSk.}$

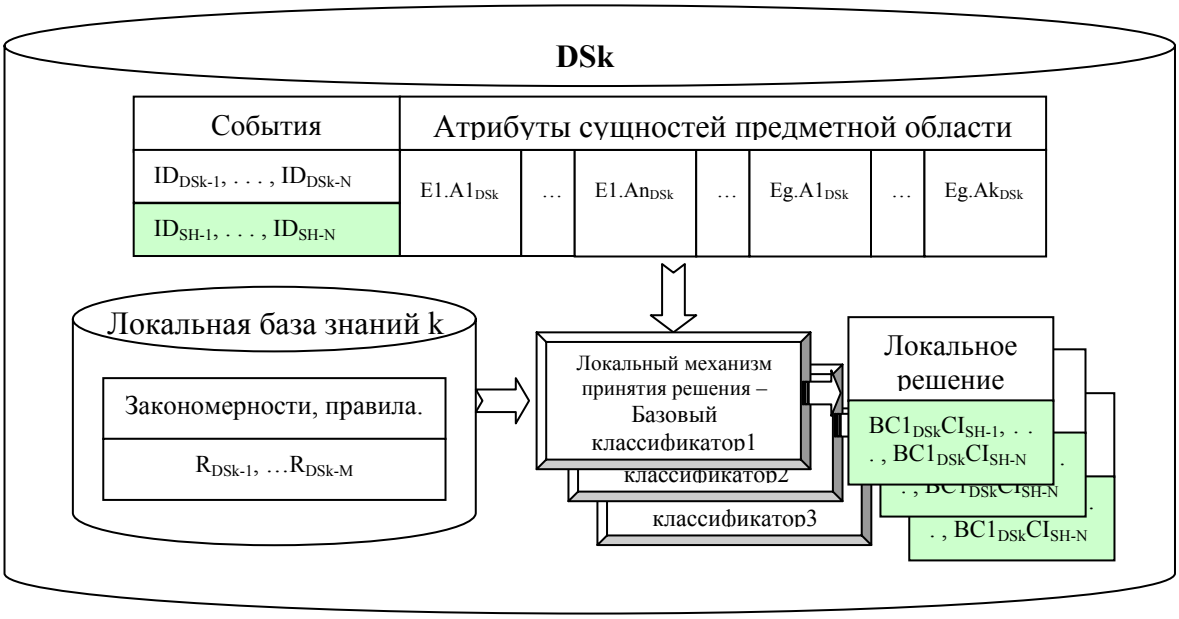

Рисунок 2 Структура объектов источника данных

Рассмотрим теперь более подробно возможные схемы объединения данных и знаний локальных источников в целях обучения на мета-уровне. Содержанием схемы объединения данных являются:

- принципы получения обучающих наборов данных метауровня;
- принципы наполнения базы знаний метауровня;
- принципы настройки механизмов принятия решений работающих с этой базой

а. Объединение закономерностей локальных баз знаний LKB_1 …LKB_N.

Этот путь возможен только при выполнении следующих условий:

- Локальные базы знаний содержат закономерности одного вида, например, правила.

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

- Атрибутивный состав {*E.A*} объединяемых источников $DS_1 \ldots DS_N$ идентичен (по крайней мере, состав атрибутов, реально используемый в закономерностях).

Только при выполнении этих условий возможно выполнение процедуры объединения закономерностей в единую базу знаний с последующей процедурой перекрестного тестирований закономерностей полученных на одних источниках – на данных других источников и определением характеристик объединяемых закономерностей.

Очевидно, что в большинстве реальных случаев данные условия не выполняются, однако имеется ряд специфичных случаев, когда именно эта схема объединения дает выигрыш по сравнению с другими схемами а именно:

- Случай когда одна и та же система принятия решений функционирует достаточно длительное время на различных разделенных источниках одинаковой структуры.
- Данные сходные по структуре находятся на различных источниках но по каким либо причинам не может быть произведено прямое объединение данных. Например, задача распознавания фальшивых транзакций в банках, в которой сами данные источника являются коммерческой тайной.

b. Прямое объединение данных $\left(ID_{SH}, \{E.A\}_{DS1} \cup \ldots \cup \{E.A\}_{DSk}\right) \mapsto \left(CI_{SH}\right)$

В этом случае фактически строится один большой набор данных, в котором каждому общему событию ID_{SH} , зафиксированному во всех объединяемых источниках данных, ставится в соответствие объединение набора атрибутов из этих источников и истинная интерпретация этого события CI_{SH}. Необходимо отметить, что при таком подходе необязательно иметь интерпретации событий в каждом из источников, достаточно, чтобы интерпретация каждого из совместно зафиксированных событий была хотя бы в одном из объединяемых источников.

Очевиден и основной недостаток этой схемы – в ней необходимо передавать и дублировать большой объем информации, что может представить большие трудности при территориально удаленных локальных источниках с обычными каналами связи.

Другие недостатки этой схемы связаны с тем, что в результате такого объединения в результирующем наборе данных, как правило, присутствуют атрибуты различных типов, что затрудняет процедуру обучения, а также затрудняется использование к процессах обучения тех данных источников, которые не вошли в число событий зафиксированных другими источниками.

c. Объединение классификаторов $\left(ID_{SH}, \{BC\_CI_{SH}\}_{DS1} \ldots \{BC\_CI_{SH}\}_{DSk}\right) \mapsto \left(CI_{SH}\right).$

В этой схеме объединяются не исходные данные локальных источников, а результаты работы уже обученных локальных механизмов принятия решений, так называемых базовых классификаторов [1]. При таком варианте объединения необходимо выполнение следующих требований:

- Наличие обученных и настроенных на каждом из источников механизмов принятия решений *{BC_CI}.*
- Наличие достаточно представительного множества базовых классификаторов для получения новых знаний на метауровне (как правило, >=3).

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

- Наличие достаточно представительного множества <*SH*> событий, зафиксированных всеми источниками. (количество событий фактически определяет объем обучающей выборки метауровня, представительность набора событий чаще всего определяется возможностями используемого метода обучения для метауровня).

Данный подход обладает рядом преимуществ, в частности:

- Существенное снижение объема информации, которой обмениваются источники и средства слияний данных, по сравнению с другими случаями (вектор атрибутов *{E.H}* в этом случае не передается);
- Возможность объединения систем, использующих разные типы данных и содержащих знания различного типа, полученных с помощью различных алгоритмов (решение проблемы *многофункциональной интеграции*) [2].
- Возможность использования всего объема данных в каждом из источников для обучения локальных механизмов принятия решения.
- Возможность использования строгих механизмов обучения объединению решений, полученных локальными механизмами принятия решений на основе схемы метаклассификации [1]. Простая модификация позволяет использовать такую схему и в многоуровневом варианте, для чего необходимо лишь включить в исходные данные результаты работы метаклассификатора предыдущего уровня.

$$\left(ID_{SH}, \{BC\_CI_{SH}\}_{DS1} \ldots \{BC\_CI_{SH}\}_{DSk}, \{MC\_CI_{SH}\}\right) \mapsto \left(CI_{SH}\right)$$

d. Объединение бинарных классификаторов.

Фактически это частный случай предыдущей схемы, когда в качестве допустимого множества значений базового или метаклассификатора используется пара альтернативных классов или логическая шкала *True/False*. Такой подход позволяет использовать единые и быстрые механизмы поиска правил и настройки механизма классификации на всех уровнях мета-классификации.

e. Объединение классификаторов и данных.

$$\left(ID_{SH}, \{E.A\}_{DS1} \ldots \{E.A\}_{DSk}, \{BC\_CI_{SH}\}_{DS1} \ldots \{BC\_CI_{SH}\}_{DSk}, \{MC\_CI_{SH}\}\right) \mapsto \left(CI_{SH}\right)$$

При таком варианте объединения каждому событию ID_{SH}, зафиксированному во всех объединяемых источниках данных, ставится в соответствие объединение каких-либо выбранных атрибутов из этих источников, результаты работы базовых механизмов вывода на локальном уровне BC_CI_{SH} или на мета-уровне MC_CI_{SH}, а также истинная интерпретация этого события CI_{SH}.

Фактически, данный вариант представляет собой комбинацию описанных выше способов объединения. Это схема, не привнося существенных преимуществ, имеет очевидные недостатки схемы прямого объединения данных. Данная схема объединения имеет ограниченную область применения и может использоваться, например, при ограниченном атрибутивном составе источников для комбинирования механизмов принятия решений, полученных с помощью различных методов над одними и теми же исходными данными.

3. Механизмы обеспечения единого представления данных различных источников

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Для решения проблем, указанных в п.2, которые возникают при объединении данных в системе принятия решений, необходим комплекс дополнительных решений, обеспечивающих практическую реализацию такой системы.

В соответствии с современными взглядами на разработку интеллектуальных информационных систем, разработку любой из них целесообразно начинать с разработки *предметной и проблемной онтологии*. Однако, если в общем случае предыдущее можно рассматривать в качестве пожелания, то для систем объединения данных иного пути просто не существует. В рамках этой онтологии должны быть определены сущности и понятия предметной области, общие для всех источников. Для каждой сущности и понятия должно быть дано исчерпывающее определение, исключающее неоднозначность трактовки на уровне локального источника. В формировании онтологии предметной области должны обязательно участвовать эксперты предметной области, работающие на уровне локальных источников данных, и эксперты предметной области, ответственные за решение задачи в целом, которые должны координировать работу экспертов локального уровня. В результате составления такой онтологии в рамках системы в целом, и в дальнейшем у всех пользователей системы складывается единое представление о предметной области и решаемой задачи.

Наличие онтологии позволяет эффективно решать ряд подзадачи задачи объединения данных. На основе единой онтологии предметной области целесообразно решать проблему идентификации сущностей, вводя для каждой сущности понятие *идентификатора сущности–ID entity*. Этот идентификатор сущности является аналогом первичного ключа для плоской таблицы. Для каждого такого идентификатора в рамках онтологии предметной области можно определить правило, которое задает способ вычисления значения этого ключа. Таким правилом, например, может являться уникальное сочетание набора нескольких атрибутов данной сущности. На уровне каждого локального источника должно быть определено правило, однозначно связывающее идентификатор сущности и локальный первичный ключ сущности в данном источнике. Это правило должно задавать:

- каким образом по значению идентификатора сущности можно получить локальный первичный ключ (по значению которого впоследствии можно получить значения всех атрибутов сущности в данном локальном источнике);
- каким образом по значению локального первичного ключа источника определить значение идентификатора сущности на уровне предметной области.

Такое правило может в частном случае просто дублировать правило определяющее идентификатор сущности (если в локальном источнике присутствуют все необходимые атрибуты сущности) или, напротив, представлять собой достаточно сложную процедуру. В вырожденном случае это может быть просто список пар «значение идентификатора сущности» - «значение локального ключа».

После того как такие правила сформированы для каждого из локальных источников на метауровне, можно легко составить перечень всех экземпляров сущностей, данные по которым находятся в локальных источниках и выделить из них те экземпляры, которые зафиксированы в нескольких источниках одновременно.

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

С помощью единой онтологии предметной области также достаточно естественно разрешаются проблемы различного представления атрибутов и различного атрибутивного состава локальных источников.

Для каждого атрибута в каждом из локальных источников вводится понятие, представляющее собой функцию интерпретации локального атрибута и имеющую тип и масштаб измерения, задаваемые в онтологии предметной области. Возможно, что для означивания атрибута будут использоваться сразу несколько атрибутов локального источника, объединенных достаточно сложными функциями. Необходимо отметить, что на уровне онтологии определяются также тип и единицы измерения атрибутов. Это позволяет на метауровне оперировать значениями атрибутов независимо от того, из какого локального источника они получены. Это понятие, очевидно, должно быть обратимо, то есть содержать функцию, которая позволяет по значению атрибута задаваемого в формате онтологии, вычислить значения атрибута в формате локального источника.

Для единообразного представления данных на различных уровнях принятия решений, с одной стороны, и для возможности использования любых данных локального источника, с другой стороны, при объединении данных целесообразно использовать схему объединения классификаторов или схему объединения бинарных классификаторов. При этом каждый локальный источник обучается на своих данных *независимо* от других локальных источников, но использует введенные в общей онтологии понятия. Такой подход позволяет также использовать и возможности схемы объединения закономерностей, так как они формируются над одним общим словарем предметной области. После того как на уровне каждого локального источника сформированы локальные классификаторы и получены спрогнозированные базовыми классификаторами значения целевого атрибута, производится процедура метаобучения по результатам принятия решения локальных классификаторов.

4. Принципы выделения данных для обучения

Для реализации процедуры обучения на локальном и метауровнях необходимо предварительно сформировать обучающие наборы данных. При формировании выборок необходимо руководствоваться двумя принципами:

a. Обучение и тестирование каждого из отдельно взятых механизмов принятия решения на каждом из уровней должно производиться на событиях, отличных от другого уровня классификации;

b. Данные обучающей и тестовой выборки каждого из уровней должны отличаться друг от друга.

Эти принципы следуют из базового положения теории обучения, согласно которому всегда возможно построить идеальный классификатор, который будет абсолютно точно работать на обучающей выборке, но давать плохие решения на любых других данных. Из этого следует, что объединение решений классификаторов на следующем уровне имеет смысл только в том случае, если результаты их работы получены на данных отличных от обучающих данных предыдущего уровня.

Рассмотрим, каким образом эти положения влияют на построение схемы классификации в системе объединения данных. Полагаем, что для каждого из локальных источников изначально зафиксировано некое подмножество

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

событий, как показано на рис. 3. Каждое такое подмножество может быть разделено на 2 части:

- События зафиксированные только в этом источнике данных *{ID_{DSg}}*;
- События зафиксированные во всех источниках данных *{ID_{SH}}.*

Как следует из вышеизложенных принципов, данные для обучения на уровне мета-классификации могут быть получены только из общей области *{ID_{SH}}.* Эти события интерпретированы подмножеством классов событий *{CI_{SH}}* общего множества возможных классов событий *{CI}.* Из этого факта следует важный вывод: построение системы принятия решения на основе распределенных источников возможно только для тех классов ситуаций, которые были зафиксированы двумя или более различными источниками.

Выбрав из альтернативного подмножества классов событий *{CI_{SH}}* те классы, для которых будет строиться система принятия решений, получаем целевое подмножество *{$CI_{SH\text{-}Task}$}.* Этому подмножеству соответствует подмножество событий *{$ID_{SH\text{-}Task}$}.* В свою очередь, это подмножество является исходным множеством событий, из которого можно создавать выборки для мета-обучения. Структура базовых и мета-классификаторов выстраивается под каждый узел дерева принятия решений системы. Поясним это более подробно.

Пусть в подмножестве *{$CI_{SH\text{-}Task}$}* определена иерархия принятия решений на классах, как это показано на рис. 4. Каждому узлу дерева принятия решений в подмножестве *{ID_{SH}-Task}* соответствует подмножество, определяемое классом события предусловия этого узла. Например, для выделенного на рис. 4 узла *DN3* это класс *Norma*, а соответствующее этому узлу подмножество событий, доступных для обучения узла–*{ID_{SH}-Task-Norma}.* Допустим, что для принятия решения на этом узле строится система поддерживающих классификаторов, показанная на рис. 4. Тогда, в соответствии с принципом "a" (см. выше в данном разделе) подмножество событий *{ID_{SH}-Task-Norma}* необходимо разделить на два подмножества – *{ID_{SH}-Task-Norma-Level1}* и *{ID_{SH}-Task-Norma-Level2}* так, чтобы в каждом из них присутствовали представители всех классов, входящих в исходы (классы результатов) узла принятия решения. Для рассматриваемого примера такими классами событий являются классы *Waiting* и *Work*. Далее, следуя принципу "b", подмножество событий каждого уровня следует разбить на обучающую и тестовую выборки, обеспечив присутствие в каждой из них событий класса *Waiting* и *Work*. Для уровня обучения базовых классификаторов *Basic Level* тестовая и обучающая выборки формируются преимущественно из подмножеств событий соответствующего источника, *не входящих* в группу общих событий *$ID_{SH.}$*. Иерархия вхождения подмножеств событий и распределение их в источнике данных для рассматриваемого примера показано на рисунке 5.

Из этого рисунка видно, насколько тесно связаны структура дерева принятия решения, структура поддерживающих классификаторов от состава и мощности обучающих данных. Из анализа рассмотренных иерархий следует важный вывод:

Проектирование структуры деревьев принятия решений и структур поддерживающих классификаторов в системах, основанных на данных распределенных источников, невозможно без предварительного анализа обучающих данных этих источников.

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

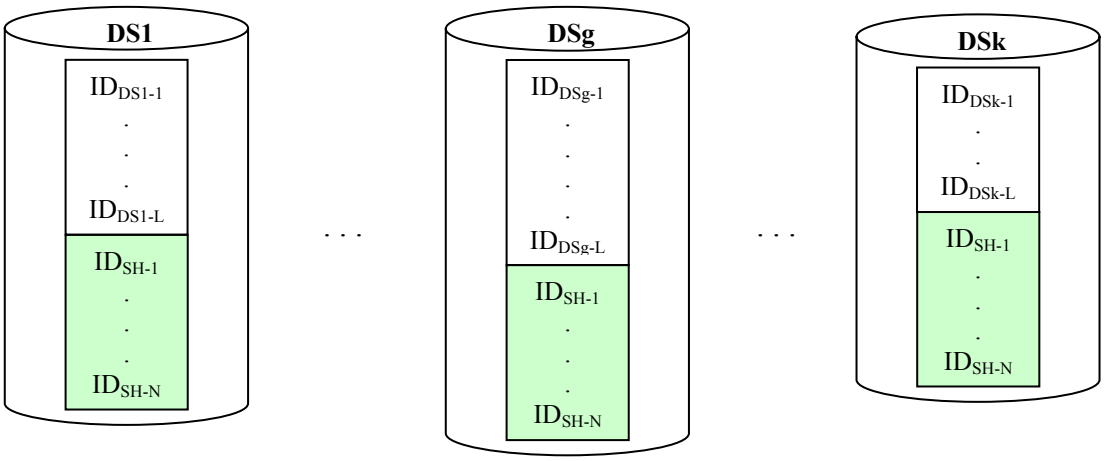

Рисунок 3 Структура распределения зафиксированных событий по источникам данных.

5. Многоагентная архитектура системы объединения данных

С учетом особенностей систем принятия решений на основе распределенных источников, отмеченных во введении, представляется целесообразным при реализации таких систем использовать архитектуру многоагентной системы. Один из вариантов архитектуры такой системы представлен на Рис. 6.

В этой архитектуре на каждом из хостов, содержащих источник данных, размещаются следующие агенты:

a. Агент формирования онтологии и подготовки данных.

С помощью этого агента на начальном этапе формирования системы производится создание и согласование общей онтологии понятий предметной области и настройка функций интерпретации на конкретную структуру данных источника. Другой функцией этого агента является выборка данных из источника по запросу.

b. Агент базовых классификаторов.

Этот агент является хранилищем базы знаний и механизмов принятия решения локального источника. Он поддерживает принятие решений базовыми классификаторами локального источника.

c. Агент обучения.

Он служит для обучения базовых классификаторов локального источника.

Особую роль играет сервер мета-обучения системы, размещенный на некотором хосте. На нем должны быть размещены следующие агенты:

a. Агент описания метаданных.

Основная функция этого агента – согласование локальных онтологий предметной области и создание единой предметной онтологии. С помощью него также производится редактирование этой онтологии, определение выражений идентификаторов сущностей и создание структур классификации.

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

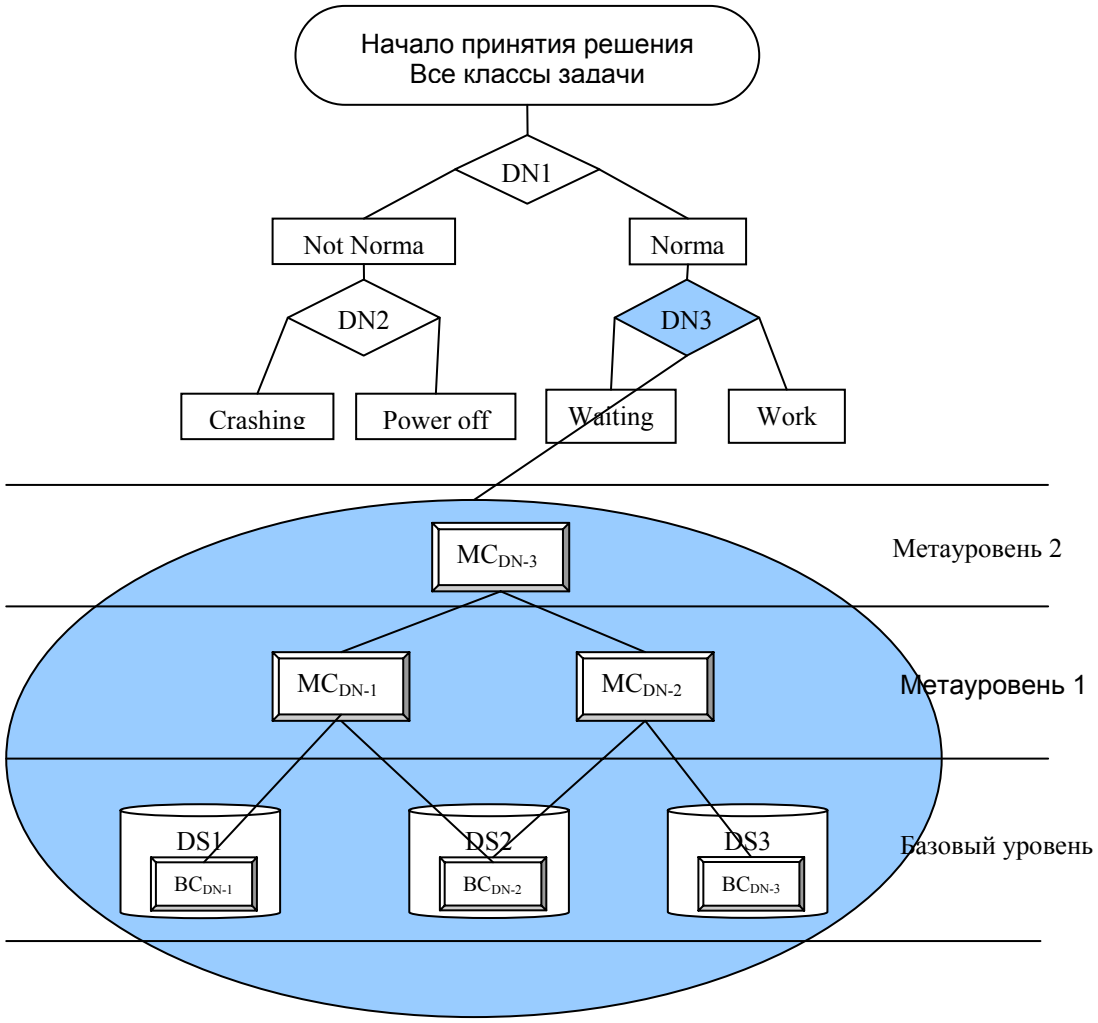

Рисунок 4 Пример иерархии принятия решений на классах и структуры поддерживающих классификаторов.

b. Агент – менеджер мета-обучения.

Функциями этого агента являются:

- Анализ данных источников и получение их метахарактеристик;
- Построение деревьев принятия решений;
- Формирование задач обучения;
- Формирование характеристик выборок;
- Формирование выборок для уровней мета-обучения.

Мета-вывод поддерживается следующими агентами:

a. *Агент мета-классификаторов*.

Этот агент является хранилищем базы знаний и механизмов принятия решения уровня мета-классификации. Он поддерживает принятие решений мета-классификаторами.

b. Агент обучения.

Он служит для обучения мета-классификаторов

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

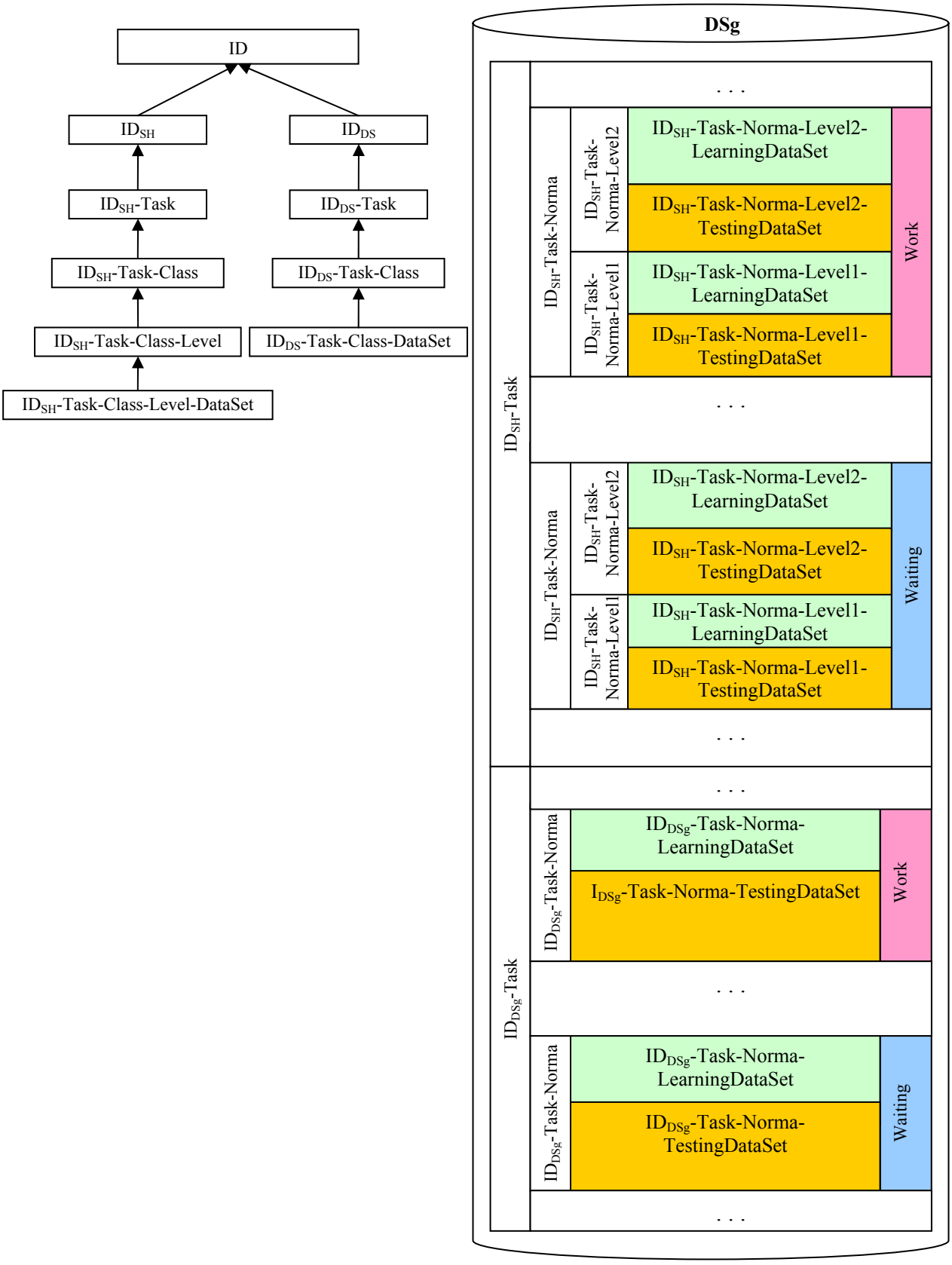

Рисунок 5 Иерархия подмножеств событий при построении структуры поддерживающих классификаторов и распределение их в источнике данных.

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

c. Агент мета-вывода.

Он служит для реализации механизма принятия решения на основе соответствующего дерева принятия решений; в режиме работы по принятию решений обобщает информацию о вновь зафиксированных событиях на локальных источниках и определяет те события, по которые могут использоваться для классификации.

6. Заключение

При разработке систем принятия решений на основе объединения данных из распределенных источников необходимо использовать общую онтологии предметной области. Из множества возможных схем объединения данных наиболее подходящей представляется схема, в которой объединяются решения классификаторов, построенных на базе локальных источников. В этом случае снимаются наиболее трудные проблемы, связанные с разнообразием свойств данных локальных источников, а также упрощается схема объединения, которая сводится к построению мета-классификатора, обучаемого на основе бинарных данных. С учетом последней рекомендации, а также в связи с распределенностью и, возможно, очень большой общей размерностью объединяемых данных, при практической реализации системы обучения целесообразно использовать многоагентную архитектуру.

Дальнейшая работа предполагает выбор и разработку конкретных методов обучения, их алгоритмизацию и программную реализацию в рамках многоагентной системы обучения объединению данных для принятия решений.

Литература

[1] *A. Prodomidis, P. Chan, S. Stolfo.* Meta-learning in distributed data mining systems: Issues and approaches, In “Advanced in Distributed Data Mining”, AAAI Press, Kargupta and Chan (eds.) `http://www.cs.columbia.edu/~sal/hpapers/DDMBOOK.ps.gz` (1999)

[2] *I. Goodman, R. Mahler, H. Nguyen*. Mathematics of Data Fusion, Kluwer academic publisher (1997)

[3] *J. Han, M. Kamber*. Data Mining. Concept and Techniques. Morgan Kaufman Publishers (2000)

[4] A Characterization of Data Mining Technologies and Processes, An Information Discovery, Inc. White Paper. `http://www.dmreview.com/whitepaper`

[5] Introduction to Data Mining and Knowledge Discovery, Third edition ©1999 by Two Crows Corporation. `http://www.twocrows.com`

[6] *M. Fayyad, G. Piatetsky-Shapiro, P. Smyth.* From Data Mining to Knowledge Discovery: An Overview. In "Advances in Knowledge Discovery and Data Mining" (Eds. U.M.Fayyad, G.Piatetsky-Shapiro, P.Smyth), Cambridge, Mass: MIT Press (1995).

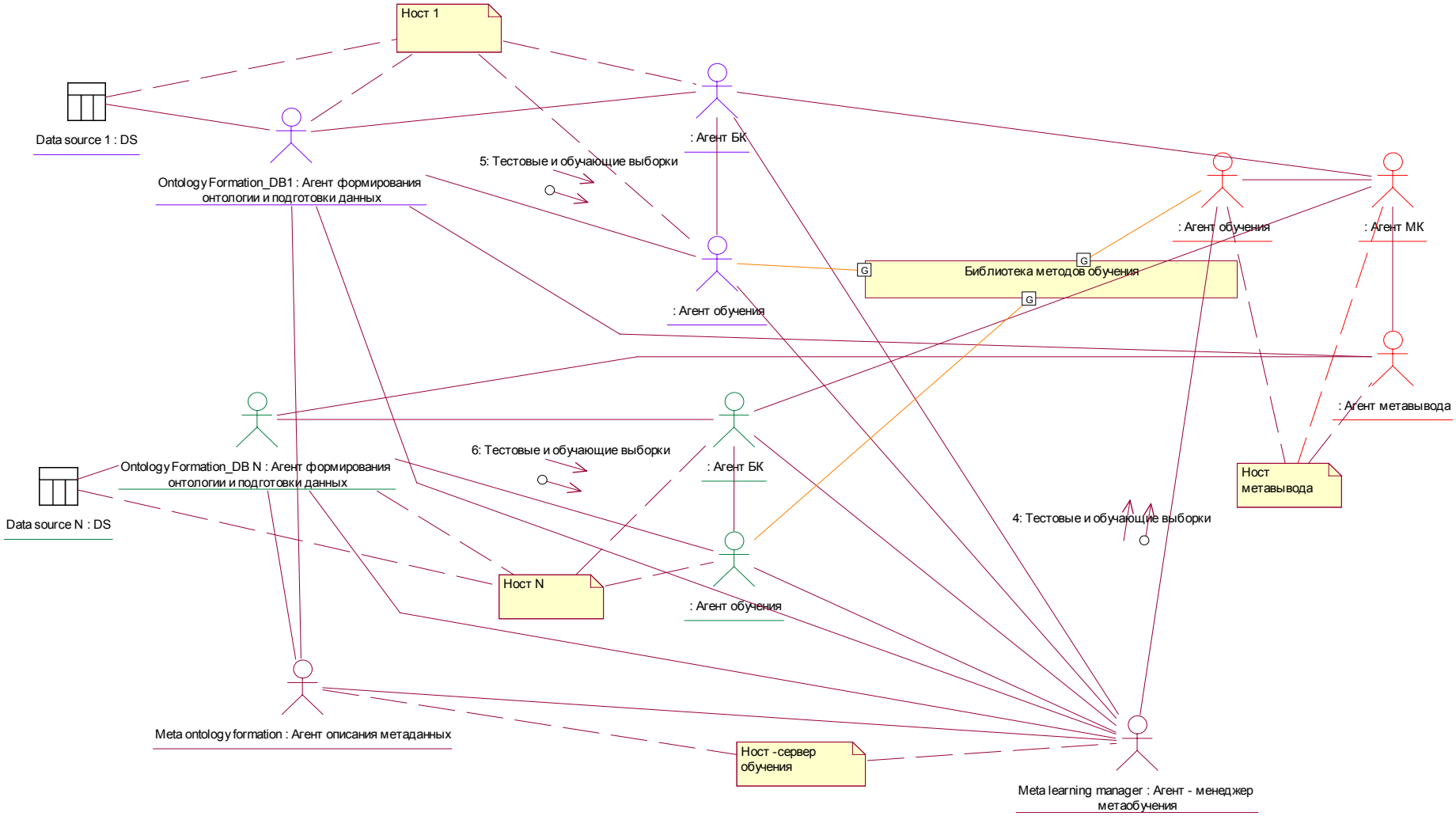

Рисунок 6. Архитектура многоагентной системы принятия решений на основе распределенных источников

Труды СПИИРАН. 2002. Вып. 1. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2002. Issue 1. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru