В. И. Городецкий, В. В. Самойлов, Ассоциативный и причинный анализ и ассоциативные байесовские сети, *Тр. СПИИРАН*, 2009, выпуск 9, 13–65

# АССОЦИАТИВНЫЙ И ПРИЧИННЫЙ АНАЛИЗ И АССОЦИАТИВНЫЕ БАЙЕСОВСКИЕ СЕТИ

ГОРОДЕЦКИЙ В.И., САМОЙЛОВ В.В.

**УДК** 681.3

*Городецкий В.И., Самойлов В.В.* **Ассоциативный и причинный анализ и ассоциативные байесовские сети.**

**Аннотация.** Поиск ассоциаций является одним из быстроразвивающихся разделов интеллектуальной обработки данных. К сожалению, традиционные подходы, развиваемые в этой области, например, при обнаружении часто встречающихся паттернов и ассоциативных правил, зачастую оказываются не в состоянии справиться с новыми приложениями, которые требуют несколько иного взгляда на методологию и технологию ассоциативного анализа. В данной работе для решения задач анализа ассоциаций привлекается неклассическая модель вероятностного пространства, которое задает класс распределений, удовлетворяющих тем ограничениям, которые накладываются доступной информацией о вероятностях некоторых, возможно, зависимых событий. В рамках этой модели, формализуемой в терминах нормированных булевых алгебр, нормированных решеток и их фрагментов, оказывается удобно решать ряд новых задач анализа ассоциаций, которые до сих пор принято относить к классу проблемных, хотя и актуальных. К их числу относятся задачи поиска редких, но сильных ассоциаций, негативных правил, а также задачи причинного анализа для принятия решений в задачах классификации. В работе предлагается единый алгоритм решения перечисленных задач, основанный на использовании структуры ассоциативной (алгебраической) сети. Этот алгоритм демонстрируется на примере.

**Ключевые слова:** ассоциативные правила, негативные правила, недоопределенное вероятностное пространство, ассоциативные байесовские сети.

*Gorodetskiy V.I.,Samoylov V.V.* **Association and Casual Rule Mining Using Associative Bayesian Networks.**

**Abstract.** Association rule search is one of the ever increasing areas of the intelligent data analysis and data mining. Unfortunately, traditional approaches of this area often are not capable to cope with new challenging problems emerging while dealing with new classes of modern applications. The latter require new viewpoint on methodology and technology of association analysis where "classical" ones fail. For efficient solution of emerging tasks of the association analysis, the paper proposes "non–classical" model of probabilistic space and its fragment called "sub–defined probabilistic space". While using algebraic view, the probabilistic models used are defined in terms of normalized Boolean algebra and lattices. Such a probabilistic model made it possible to cope with several challenging association analysis tasks. Between them, the proposed algorithm is capable of search for rare but "strong" association rules, mining negative rules of any forms and mining cause consequence rules. All these tasks are solved within the same framework called associative (algebraic) Bayesian network. The basic algorithm is demonstrated by simple case study, although the algorithm and corresponding software developed for this purpose were validated on an application of real life scale.

**Keywords:** association rules, negative rules, sub defined probabilistic space, associative Bayesian networks.

**1. Введение.** Поиск ассоциаций является одним из быстроразвивающихся разделов интеллектуальной обработки данных. Исследования в этой области, которые насчитывают уже почти 20 летнюю историю, носят главным образом прикладной характер, поскольку они с самого начала инициированы практическими потребностями и фактически до настоящего времени обслуживают потребности, которые возникают на практике. Традиционно этот подход используется в качестве теоретического базиса в области маркетинговых исследований при изучении и прогнозировании спроса на товары в различных регионах. Он используется также при анализе сезонных трендов покупательского спроса, вариантов покупательских корзин, что необходимо для улучшения организации торговли внутри конкретного магазина и др. В настоящее время ассоциативный анализ шагнул далеко за пределы чисто маркетинговых исследований. Он находит все более широкое применение при изучении и прогнозировании общественного мнения в социологии, политологии и других гуманитарных областях, где особенности статистической информации, которой приходится оперировать, создают значительные трудности для использования других, более "интеллектуальных" методов и подходов. В последнее время к ассоциативному анализу проявляется интерес специалисты, исследующие различные Интернет сообщества, которые спонтанно возникают в форме социальных сетей, форумов, блогов и т. п.

К сожалению, традиционные подходы, развиваемые в области анализа ассоциаций, например при обнаружении часто встречающихся паттернов и ассоциативных правил, зачастую не в состоянии справиться с новыми приложениями, которые требуют несколько иного взгляда на методологию и технологию ассоциативного анализа.

Одно из принципиальных отличий современных задач от "классических" состоит в том, что используемые данные обычно представляются в *реляционной базе данных* достаточно сложной структуры, которая включает в себя десятки и более различных таблиц[1]. Этот факт создает определенные проблемы, делая методы анализа ассоциаций более громоздкими и менее эффективными. Однако с этой проблемой "классические" методы в основном справляются [14, 19].

[1] В "сыром виде" данные чаще представляются текстовыми файлами, которые необходимо декодировать и преобразовать к реляционной форме.

Другая особенность современных систем, основанных на знаниях, состоит в том, что они, как правило, на мета уровне используют онтологию. Это дополнительно осложняет проблему анализа ассоциаций, поскольку влечет необходимость учета не только реляционной структуры данных, но также и *отношения* на множестве понятий предметной области, представленные в *онтологии.* Более того, часто требуется, чтобы и результаты анализа ассоциаций были представлены в терминах онтологии.

По прежнему проблемными остаются некоторые задачи ассоциативного анализа даже при традиционной модели данных, и при этом речь идет не только о повышении эффективности алгоритмов ассоциативного анализа в крупномасштабных задачах. К сожалению, современный ассоциативный анализ базируется на ряде не вполне обоснованных предположений. Например, предположение о том, что ассоциативные правила следует искать только на множестве часто встречающихся паттернов, используемое в большинстве известных алгоритмов, автоматически исключает возможность найти редко встречающиеся, но очень сильные ассоциации. В ряде приложений именно они являются предметом наибольшего интереса. Примером является криминалистика, анализ Интернет–сообществ на предмет обнаружения социально опасных сообществ и др. Хотя задача поиска редких, но сильных ассоциаций и не остается без внимания исследователей, тем не менее, пока отсутствуют хорошие алгоритмы их поиска. Другое не вполне обоснованное предположение, используемое в большинстве работ, состоит в том, что рассматриваются ассоциативные правила с посылками и заключениями в виде конъюнкций положительно означенных пропозиций. Такие варианты ассоциаций искать гораздо легче [31], однако при этом теряется очень много полезных закономерностей. Например при этом теряются все закономерности, которые принято называть *запретами*, которым соответствуют ассоциации, содержащие в заключении отрицательную литеру. В то же время они могут содержать достаточно "сильные" закономерности. И хотя в последнее время появились работы, исследующие проблему поиска таких ассоциаций правил, однако "хороших" решений этой задачи пока не предложено. Имеются и другие недостатки известных алгоритмов поиска ассоциаций.

Одной из причин такого состояния исследований в области ассоциативного анализа является слабый акцент на вероятностную интерпретацию моделей и результатов ассоциативного анализа.

Большую пользу может дать привлечение неклассических моделей теории вероятностей, когда не вводится понятие пространства элементарных событий, а сама модель определяется через множество зависимых событий [15, 16, 18]. Такая модель включает в себя гораздо более широкий класс вероятностных пространств, чем это предполагается в "классической" модели. Информация, которая обычно доступна в практических задачах, является весьма ограниченной, включает в себя множество зависимых событий, не содержащих информацию, необходимую для однозначного задания распределения вероятностей. Возможно, именно по этой причине по большей части практически ориентированные исследования в области анализа ассоциаций относительно слабо используют вероятностную трактовку компонент проблемы, изредка привлекая вероятностные свойства компонент задачи поиска для сокращения перебора. Тем самым возможности ассоциативного анализа сильно обедняются. Иногда это приводит и к некорректной терминологии[1].

В данной работе для решения задач анализа ассоциаций используется вероятностная трактовка постановок задач ассоциативного анализа, и для их решения привлекается неклассическая модель вероятностного пространства, которое является недоопределенным в том смысле, что оно задает класс распределений, удовлетворяющих тем ограничениям, которые накладываются доступной информацией о вероятностях некоторых, возможно, зависимых событий. В рамках этой модели, формализуемой в терминах нормированных булевых алгебр, нормированных решеток и их фрагментов, оказывается удобным решать ряд новых задач анализа ассоциаций, которые до сих пор принято относить к классу проблемных, хотя и актуальных. К их числу относятся задачи поиска редких, но сильных ассоциаций, негативных правил, а также задачи причинного анализа и извлечения правила принятия решений в задачах классификации.

Дальнейший материал данной работы организован следующим образом.

В разделе 2 приводится краткий анализ современного состояния исследований (по работам до 2008 г. включительно) в области анализа

[1] Одним из примеров некорректности является само понятие ассоциативного правила, которое везде трактуется как импликация, хотя на самом деле оно импликацией не является. Это показано в дальнейшем материале работы.

ассоциаций. При этом большое внимание уделяется работам, затрагивающим наиболее проблемные задачи.

В разделе 3 вводится модель вероятностного пространства, использующая в своей основе алгебраический подход, а также понятие недоопределенного вероятностного пространства, которое в менее общей форме, предложено одним из авторов в начале 1990-х гг. [2, 3, 17] и более детально описано позже в [5, 6] и др.

В разделе 4 дается краткое описание модели Алгебраической Байесовской Сети (АБС), которая ранее описана в названных работах, а также в работе [4]. В настоящей работе эта модель называется Ассоциативной Байесовской Сетью, хотя имеется в виду та же самая модель АБС. Этот термин используется здесь для того, чтобы подчеркнуть прагматику использования этой модели в данной работе, в которой решается задача поиска и анализа ассоциаций.

Раздел 5 содержит основные новые результаты работы. В нем описываются модели и алгоритмы поиска ассоциативных правил в модели, которая названа моделью типа *уверенность-зависимость-причинность*, чтобы отличить ее от двух основных используемых в настоящее время моделей, одна из которых называется моделью типа *поддержка-уверенность* а вторая – моделью типа *поддержка-уверенность-зависимость* [31]. В этом разделе показывается, каким образом модель АБС может быть эффективно использована для решения ряда проблемных задач поиска ассоциаций, в частности, соответствующих причинно–следственным связям.

В разделе 6 кратко рассматривается алгоритм предобработки данных для перехода от модели данных к модели АБС в терминах агрегированных признаков. Кроме того, здесь же основные результаты работы поясняются и демонстрируются на примере конкретного приложения, в котором решается задача обучения предсказанию предпочтений зрителей при выборе фильмов.

В заключении формулируются основные результаты работы и определяется область их приложений.

**2. Модели ассоциативных правил и алгоритмы их генерации: Краткий обзор.** Ассоциативные правила дают полезную новую информацию о множестве данных в целом и активно используются в практике. В современной научной литературе этот раздел извлечения знаний из данных представлен достаточно широко (например, [9, 20, 31] и многие другие). Имеется ряд коммерческих инструментальных систем, ориентированных на извлечение ассоциативных правил.

Рассмотрим кратко варианты постановок задачи поиска ассоциативных правил, основные предположения и ограничения, а также опишем основные идеи известных методов их поиска с акцентом на их достоинства и недостатки. Пусть $\boldsymbol{A}$ есть некоторое множество различных атрибутов (иногда говорят "объектов", "элементов"), а данные обучения представлены множеством $\boldsymbol{D}$, в котором представлены записи некоторого реального процесса. Без потери общности предположим, что атрибуты множества $\boldsymbol{A}$ могут быть упорядочены, например, лексикографически. Хотя это упорядочение может быть искусственным, оно иногда удобно для построения более понятных, а иногда и более эффективных алгоритмов извлечения ассоциативных правил из данных. Поэтому вместо термина "набор" элементов иногда используется термин "последовательность".

**2.1. Модель поддержка–уверенность.** Введем понятие ассоциативных правил [9] и рассмотрим модель, которую принято называть моделью типа "*поддержка–уверенность*" (*support–confidence framework"* [31]). Пусть $\boldsymbol{D}$ – база данных (транзакций), $\boldsymbol{A}$ – множество символов, которые используются для обозначения объектов, $X \in \boldsymbol{A}$ – подмножество (подпоследовательность, паттерн) символов из множества, и $\boldsymbol{\Lambda}(X)$ – подмножество множества транзакций из множества $\boldsymbol{D}$, которые содержат набор $X$ в качестве подмножества. Для характеристики свойств набора $X$ в базе данных используют отношение мощности множества $\boldsymbol{\Lambda}(X)$ к мощности множества $\boldsymbol{D}$. Эту величину принято называть *поддержкой* (*support*) последовательности $X$ в множестве $\boldsymbol{D}$:

$$supp(X) = |\boldsymbol{\Lambda}(X)|/|\boldsymbol{D}|. \quad (1)$$

Пусть даны два набора элементов $X \in \boldsymbol{A}$ и $Y \in \boldsymbol{A}$, причем $X$ и $Y$ не имеют общих элементов, и пусть $\sigma$ и $\gamma$ – числа из интервала [*0, 1*]. Величину $\gamma$ будем называть *порогом для меры уверенности*, а величину $\sigma$ – *порогом для поддержки* ассоциативного правила.

*Определение 1.* Говорят [9], что выражение вида $X \to Y$ есть *ассоциативное правило с порогом уверенности $\gamma$ и порогом поддержки* $\sigma$ ($\sigma$, $\gamma$ – ассоциативное правило), если

$$1.\ |\boldsymbol{\Lambda}(X + Y)|/|\boldsymbol{D}| \geq \sigma \quad (2)$$

и

$$2.\ |\boldsymbol{\Lambda}(X+Y)|/|\boldsymbol{\Lambda}(X)| \geq \gamma \qquad (3)$$

В формулах (2), (3) $X$ и $Y$ являются наборами имен элементов из $\boldsymbol{A}$, а $X+Y$ есть объединение их множеств имен. Поэтому величина в левой части формулы (2) есть значение поддержки для $X+Y$ в множестве данных $\boldsymbol{D}$ (см. формулу (1)), а величина в левой части формулы (3) есть значение поддержки для подмножества $X+Y$ в выборке данных $\boldsymbol{\Lambda}(X)$. Величину

$$|\boldsymbol{\Lambda}(X+Y)|/|\boldsymbol{\Lambda}(X)|$$

принято называть *мерой уверенности* для ассоциативного правила $X \to Y$ в выборке $\boldsymbol{\Lambda}$. Обычно ее обозначают символом *conf*( $\boldsymbol{\Lambda}, X \to Y$ ) (от английского термина "*confidence*"). Далее, когда по тексту ясно, о какой выборке идет речь, символ $\boldsymbol{\Lambda}$ в выражении для меры уверенности будет опускаться.

Введение символа "+" для обозначения объединения операции объединения множества имен при анализе ассоциаций вызвана необходимостью различать ее с операцией "$\cup$", которая далее используется в алгебре множеств (алгебре событий) для операций со множествами. $X$ и $Y$ являются множествами непересекающихся множеств символов–имен, поэтому для них теоретико–множественная операция "$\cap$", если они трактуются просто как множества символов, приводила бы к пустому множеству $X \cap Y = \varnothing$. Однако в примерах базы данных $\boldsymbol{D}$ наборы $X$ и $Y$ могут встречаться вместе. В ней каждая из величин $X$ и $Y$ имеет смысл сложных случайных событий, которые происходят при *совместном* появлении более простых случайных событий, соответствующих символам наборов $X$ или $Y$, и как случайные события они могут быть зависимыми. Поэтому совместная вероятность $p(X \cap Y)$ в алгебре событий есть вероятность сложного события, обозначенного выше как $X+Y$ в множестве имен объектов[1].

С учетом содержательной интерпретации принятых обозначений величинам $\sigma$ и $\gamma$ может быть дана простая вероятностная трактовка. В соответствии с определениями (2) и (3), первая из них является

[1] Имеется другой способ разрешения такой коллизии обозначений. Для этого некоторые авторы предлагают множество $X$, рассматриваемое как случайное событие, обозначать иначе, например, $\hat{X}$. Однако это делает громоздким обозначение.

оценкой вероятности появления в выборке $\boldsymbol{A}$ набора элементов (паттерна) $X + Y$ , т. е. паттерна, содержащего одновременно оба набора $X$ и $Y$. Поэтому эта величина соответствует оценке вероятности *совместного* появления паттернов $X$ и $Y$ в выборке $\boldsymbol{D}$:

$$p\,(\,X \cap Y\,) = |\,\boldsymbol{A}(X+Y)\,|/|\boldsymbol{D}| = supp\,(\,X+Y\,). \qquad (4)$$

Вторая величина является оценкой условной вероятности появления паттерна $Y$ в выборке при условии, что паттерн $X$ в нем присутствует:

$$p\,(Y/X) = |\,\boldsymbol{A}(X+Y)\,|/|\,\boldsymbol{A}(X)\,| = conf\,(X \to Y). \qquad (5)$$

В зарубежной литературе выражение $X \to Y$ обычно называется импликацией. Однако это по меньшей мере неточно. Различие между выражением $X \to Y$ становится очевидным, если рассматривать их как события в вероятностном пространстве, а именно этот аспект прежде всего интересен при анализе ассоциаций. Действительно,

$$p(\,X \to Y\,) = p\,(\,Y/X\,) = p\,(\,X \cap Y\,)/\,p\,(\,X\,),$$

и эта трактовка будет далее использоваться при поиске ассоциативных правил. Что касается вероятностной трактовки импликации, то

$$p\,(\,X \supset Y\,) = p\,(X \cap \overline{Y}\,) = 1 - p(Y) + p(X \cap Y).$$

Итак, задача поиска ассоциативных правил в модели *поддержка–уверенность* формулируется таким образом [9]:

*Для заданных значений порогов поддержки* σ *и меры уверенности* γ *и для заданной базы обучающих данных* $\boldsymbol{D}$ *найти все такие пары паттернов* $X$ *и* $Y$, *для которых отношение* $X \to Y$ *является* (σ, γ) - *ассоциативным правилом.*

В этой постановке любая процедура поиска ассоциативных правил в модели *поддержка–уверенность* укладывается в рамки следующей обобщенной схемы [9]:

1). Найти все часто встречающиеся паттерны $A_k \in \boldsymbol{A}$ с поддержкой не менее чем σ, т. е.

$$supp(\,A_k\,) \geq \sigma; \qquad (6)$$

2.) Среди всех часто встречающихся паттернов $A_k \in \boldsymbol{A}$ , найденных на шаге 1, сгенерировать ассоциативные правила $X_i \to Y_j$, такие, что:

1). $X_i + Y_j = A_k$,

т. е. часто встречающийся паттерн $A_k$ состоит из символов, входящих в множества $X_i$ и $Y_j = A_k \setminus X_i$,

2). $conf(X_i \rightarrow Y_j) \geq \gamma.$ (7)

Заметим, что параметры σ и γ, задаваемые экспертами, являются параметрами алгоритма, и они могут использоваться для поиска баланса между числом сгенерированных правил (чем больше σ и γ, тем меньше правил) и качеством правил (чем больше σ и γ, тем более "сильные" правила будут сгенерированы).

Большинство разработанных к настоящему времени алгоритмов поиска ассоциативных правил, следующих модели *поддержка–уверенность*, имеют переборный характер. Их различия главным образом состоят в использовании ряда разных приемов, позволяющих ускорить поиск. Основу большинства известных в настоящее время алгоритмов поиска ассоциативных правил составляет алгоритм, известный под названием Apriori, предложенный в работе [10], хотя несколько алгоритмов предложены ранее, например алгоритм AIS [11]. Дадим краткое описание идеи алгоритма Apriori, следуя [20].

Пусть экспертом заданы пороговые значения σ и γ. Основная идея алгоритма базируется на свойстве *антимонотонности* функции *supp*, в соответствии с которым для двух паттернов $A_k \in \boldsymbol{A}$ и $A_r \in \boldsymbol{A}$, таких, что $A_k \subset A_r$, справедливо

$$supp(A_k) \geq supp(A_r).$$

Действительно, условие $A_k \subset A_r$ означает, что менее мощное подмножество, составляющее паттерн $A_k$, обязательно входит во все те транзакции, в которые входит паттерн $A_r$. Однако паттерн $A_k$ может входить также и в другие транзакции базы данных, что и влечет за собой отмеченное свойство. При вероятностной трактовке меры *supp* это свойство выводится из свойств вероятностей. Поэтому, если установлено, что $supp(A_k) < \sigma$, то поддержка любого другого паттерна, который получается из $A_k$ путем добавления хотя бы одного символа, будет также иметь значение поддержки менее σ. Поэтому такие цепочки можно исключить из генерации ассоциативных правил, что позволяет отсекать неперспективные ветви поиска.

Другие приемы для сокращения поиска связаны с использованием хэш-функций, уменьшением числа сканируемых транзакций в базе обучающих данных, разбиением баз данных с поиском частых последовательностей по подвыборке из базы транзакций и рядом других [9], [20], [31].

Однако существуют и более эффективные алгоритмы, которые не связаны с многократным просмотром базы данных, что необходимо для генерации кандидатов в множества часто встречающихся паттернов, задаваемых формулой (6) в алгоритмах на основе алгоритма Apriori. Один из них, называемый *методом возрастающих паттернов* (*Frequent Pattern growth, FP-growth*) [21], базируется на:

1) построении дерева последовательностей возрастающей длины (*FP-tree*);

2) на последующем извлечении из него последовательностей, для которых значение функции *supp* не меньше заданного порога.

Поскольку этот метод интересен и в настоящее время наиболее эффективен, дадим его краткое описание, следуя работе [20].

*Алгоритм 1. Построение дерева паттернов возрастающей длины.* Все последовательности, которые потенциально могут быть кандидатами в множество часто встречающихся паттернов, представляются в виде дерева, которое строится с помощью достаточно простых действий, описываемых далее. Эти действия поясняются на рис. 1, на котором демонстрируется построение этого дерева для транзакционной базы данных, представленной на рис. 2.

1. Находятся все одноэлементные последовательности, удовлетворяющие заданному порогу σ. Они упорядочиваются в соответствии с приписанными им значениями функции *supp* и записываются в виде столбца таблицы в порядке убывания функции $|\Lambda(X_i^{(1)})|$, где $X_i^{(1)}$ – одноэлементные последовательности, удовлетворяющие условию $supp(X_i^{(1)}) \geq \sigma$), значения которой записываются во втором столбце этой таблицы (рис.1). Обозначим принятый порядок символом $L$, а само

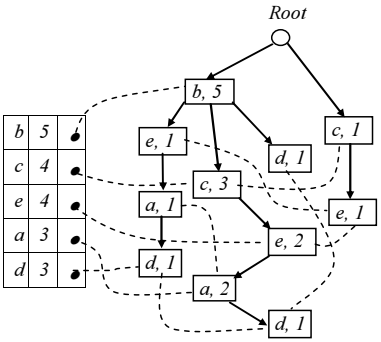


Рис. 1. Дерево последовательностей для транзакционной базы данных, представленной на рис. 2.

множество – символом $A(\sigma)$.

2. Вводится корневой узел *Root* дерева последовательностей; обозначим его символом *T*.

3. Сканируются транзакции базы данных и для каждой транзакции (см. рис. 2) проделывается следующее:

3.1. Из транзакции выбирается и сортируется в соответствии с некоторым выбранным порядком *L* (например, лексикографическим) последовательность максимальной длины, которая состоит из символов множества $A(\sigma)$, представленных в первом столбце таблицы (рис. 1). Пусть *Trans* – очередная обрабатываемая транзакция. Обозначим символом *X* упорядоченную последовательность символов транзакции *Trans* . Пусть $X = x\,Y$ , где $x$ – первый элемент (префикс) последовательности $X$ , а $Y$ – оставшаяся ее часть. Заметим, что первым символом последовательности является тот, который по порядку *L* является "наибольшим".

3.2. Если к текущему шагу дерево *T* уже имеет узел, помеченный символом $x$, который непосредственно следует за корнем *Root*, то к счетчику, поставленному в соответствие этому узлу, прибавляется 1. В противном случае создается новый узел – потомок корня *Root* с таким же именем, и его счетчику присваивается значение 1. Назовем эту процедуру *InsertTree(*$x\,Y$*, N),* где *N* = *Root.* Заметим, что символом *N* здесь обозначается родительский узел в дереве *T*, за которым следует узел, помеченный символом первого элемента последовательности. Он в этой процедуре играет роль первого аргумента.

| *r* | | Транзакция |
|---|---|---|
| 1 | → | *abde* |
| 2 | → | *bd* |
| 3 | → | *ce* |
| 4 | → | *bc* |
| 5 | → | *abce* |
| 6 | → | *abcde* |

Рис. 2. Пример транзакционной базы данных

Далее рекурсивно для всех символов, составляющих оставшуюся часть последовательности $Y$ , выполняются действия, описанные в п. 3.3.

3.3. Если $Y$ непусто, то используется рекурсивно процедура *InsertTree(*$x\,Y$*, N)* при *N* = $x$, где $x$ – последний из обработанных символов. Если $Y$ пусто, то перейти к п. 3.4.

3.4. Если множество нерассмотренных транзакций непусто, то перейти к обработке следующей транзакции (см. п. 3.1), иначе – конец алгоритма.

Результат работы алгоритма 1 – это дерево последователь-ностей *T*, соответствующее обрабатываемой транзакцион-ной базе данных и заданному порогу σ для функции *supp*. Построенное дерево представляется в некотором формате, например, в виде структуры данных ***B***-дерева [13]. На рис. 1. показано дерево последовательностей, которое получено применительно к транзакционной базе данных, представленной на рис. 2.

*Алгоритм 2. Извлечение* σ - *последовательностей из дерева T* предназначен для извлечения частых последовательностей из дерева *T*. Он состоит их трех шагов. На первом шаге строится так называемая "условная база последовательностей", в которую входят префиксы последовательностей, которые в качестве *последнего* символа имеют один из символов множества $A(\sigma)$. Построение такой базы производится для всех $\alpha \in A(\sigma)$ за исключением первого символа (по введенному порядку *L*), причем поиск этих префиксов выполняется в порядке, обратном порядку *L*. На втором шаге для каждой такой условной базы последовательностей строится условное дерево последовательностей, для чего используется алгоритм 1. Далее из него достаточно просто уже извлекаются все σ - последовательности.

Аргументами алгоритма являются *T*–дерево последовательностей, построенное *алгоритмом 1*, и α – один из символов множества $\alpha \in A(\sigma)$, кроме первого по порядку *L*. В основе этого алгоритма лежит процедура FP-growth$(T, \alpha)$, $\alpha \in A(\sigma)$ [20]. Множество $A(\sigma)$ заносится в первый столбец таблицы, которая формируется в процессе *алгоритма 2*, при этом они записываются в него в порядке возрастания значений поддержки (см. таблицу на рис. 1), т. е. в порядке, обратном порядку *L*.

1. Из множества $A(\sigma)$ выбирается символ α, обладающий наименьшей поддержкой в базе транзакций[1], и в дереве *T* выбираются все пути, ведущие из корня к узлам с данным именем.

2. Для каждого из таких путей решается задача, описываемая далее в п. 3.

3. Пусть рассматривается путь *P*. Заметим, что в этот путь узел α не входит.

[1] Напомним, что символы $\alpha \in A(\sigma)$ являются σ – последовательностями длины 1.

3.1. Если $P$ – простой путь[1], то для него формируется множество всех возможных последовательностей из символов, встречающихся в нем. Обозначим это множество последовательностей символом $\boldsymbol{B}$, а произвольную последовательность в ней – символом $\beta$, $\beta \in \boldsymbol{B}$. Далее генерируется множество последовательностей вида $\beta \cup \alpha$ с поддержкой *supp*( $\beta \cup \alpha$ ), равной минимальному значению поддержки на множестве узлов, входящих в $\beta \cup \alpha$. $\sigma$ - последовательности из множества $\boldsymbol{B}$ вместе со значениями поддержки записываются в столбец 4 Табл. 1. в строку, отвечающую символу $\alpha$.

3.2. Если путь $P$ не является простым, то для каждого из путей в узел с именем $\alpha$ формируется префикс $\alpha_i$ узла $\alpha$, $\beta = \alpha_i \cup \alpha$, и каждой из последовательностей $\beta$ присваивается значение поддержки, равное минимальной поддержке узла в нем.

4. Формирование условной базы последовательностей, содержащей все последовательности $\beta$, найденные в п.3.2, с присвоенными им значениями поддержки *supp*. Условная база последовательностей записывается в столбец 2 в строку, отвечающую символу $\alpha$ (см. Табл. 1.).

Табл. 1. **Результаты работы алгоритма 2**

| Символ $\alpha. \in A(\sigma)$ | Условная база данных последовательностей | Условное FP-дерево | Множество $\sigma$-последовательностей |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| d | (bea: 1), (bcea: 1), (b: 1) | (bea: 2), (b: 1) | bd |
| a | (be: 1), (bce: 2) | (be: 3) | ba, be, ea, bea |
| e | (b: 1), (bc: 2), (c:1) | (b: 3), (c: 3) | be, ce |
| c | (b: 3) | (b:3) | bc |

5. Формирование условного дерева последовательностей в соответствии с алгоритмом 1 для условной базы последовательностей, которая состоит из всех префиксов символа $\alpha$. Пути условного дерева, из которого исключаются узлы, имеющие

[1] Имеется в виду путь, не содержащий ответвлений.

поддержку, меньшую σ, записываются в столбец 3 в строку, отвечающую символу α. Удаление из исходного дерева узлов, содержащих символ α.. Удаление символа α. из множества $\boldsymbol{A}(\sigma)$.

Заметим, что поскольку α. соответствует символу с наименьшим значением поддержки в множестве $\boldsymbol{A}(\sigma)$, то на шаге 5 всегда удаляются некоторые *листья* текущего дерева *T*.

6. Если множество $\boldsymbol{A}(\sigma)$ непустое (дерево *T* пусто, если оно не содержит узлов, кроме узла *Root*), то процедура повторяется, начиная с п. 1, для следующего символа в первом столбце таблицы (см. Табл. 1).

7. По условным деревьям генерируется множество всех σ - последовательностей в каждой из строк. σ - последовательности вместе со значениями поддержки записываются в столбец 4 таблицы.

**2.2. Модель *поддержка–уверенность–зависимость*.** Понятие ассоциативного правила, введенное условиями (2) и (3), обладает большим недостатком, поскольку не учитывает возможную вероятностную независимость паттернов $X$ и $Y$ при высоком значении поддержки. В этом случае говорить об ассоциативной связи этих паттернов бессмысленно. Высокое значение поддержки при независимости паттернов может иметь место просто за счет больших значений вероятностей самих паттернов, когда вероятностная связь между ними отсутствует:

$$\begin{gathered}\text{если } p(X \cap Y) = p(X)\,p(Y), \text{ то} \\ supp(X+Y) \approx supp(X) \times supp(Y) \geq \sigma,\end{gathered} \tag{8}$$

хотя $p(Y/X) = p(Y)$ и $p(X/Y) = p(X)$, т. е. ассоциативная связь между паттернами $X$ и $Y$ отсутствует. Поэтому поиск ассоциаций в таком случае не представляет интереса [25], что игнорировалось ранее моделью *поддержка–уверенность*, которая долгие годы была и остается сейчас наиболее популярной.

В отличие от классического подхода, рассматриваемого в модели *поддержка–уверенность*, в модели, описанной в работе [25], предложена дополнительная мера качества ассоциативного правила, которая вводит еще один параметр для выбора или отклонения правила. Она использует хорошо известную статистическую меру зависимости между паттернами:

$$I = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{9}$$

Близость этой величины к единице свидетельствует о слабой статистической зависимости между паттернами $X$ и $Y$. В модели Г. Пятецкого-Шапиро она используется в качестве пороговой характеристики:

$$|\frac{P(X \cap Y)}{P(X)P(Y)} - 1| \geq \delta_{min}, \quad (10)$$

где величина $\delta_{min}$ названа автором "минимальный интерес".

Соответственно, в данной модели определение ассоциативного правило расширено требованием (10). Оно дается таким образом.

*Определение 2.* Ассоциативным правилом называется отношение $X \to Y$, заданное на паре паттернов $X$ и $Y$, для которых выполнены следующие условия:

1). Множества $X$ и $Y$ не содержат общих элементов, $X + Y = \varnothing$;

2). $p(X \cap Y) \geq \sigma_{min}$;

3). $p(Y / X) \geq \gamma_{min}$;

4). $|\frac{P(X \cap Y)}{P(X)P(Y)} - 1| \geq \delta_{min}$.

Параметрами алгоритмов для поиска ассоциативных правил, удовлетворяющих определению 2, являются значения минимальной поддержки $\sigma_{min}$, минимальной уверенности $\gamma_{min}$ и минимального интереса $\delta_{min}$. Алгоритмы поиска ассоциативных правил в данном случае отличаются от алгоритмов типа Apriori дополнительной проверкой условия 4 и отсечением правил, которые ему не удовлетворяют.

**2.3. Другие модели.** Существует еще несколько моделей ассоциативных правил и алгоритмов их поиска, которые, как и модель *поддержка–уверенность–зависимость*, акцентируют внимание на обязательном наличии статистической зависимости между паттернами посылки и заключения правила.

В модели [12] для оценки зависимости между компонентами паттерна используется $\chi^2$ тест, - классический критерий математической статистики. Этот тест проверяет гипотезу о независимости компонент паттерна. При этом используются известные выражения для независимых событий, выражаемые в терминах паттернов следующим образом:

$$p(X \cap Y) = p(X)\,p(Y),$$
$$p(X \cap Y \cap Z) = p(X)\,p(Y)\,p(Z), \qquad (11)$$

Алгоритм проверки гипотезы состоит:

1) в вычислении оценки вероятностей отдельных паттернов по выборке;

2) вычислении оценки вероятности их совместного появления в выборке;

3) оценке значимости различия между этими величинами по критерию $\chi^2$ для заданного порога отсечки, которое соответствует заданному уровню доверия, например, 0,95.

Однако известно, что критерий $\chi^2$ обладает низкой точностью в том случае, когда оценивается достоверность зависимости при малых значениях вероятностей, входящих в формулы (10), что ограничивает его применение. В любом случае он неприменим при поиске редко встречающихся, но сильных зависимостей.

**2.4. Негативные ассоциативные правила.** Ассоциативные правила, рассмотренные ранее, принято называть *позитивными* (*positive*) правилами потому, что они в посылке и заключении содержат паттерны, утверждающие присутствие элементов утверждения о присутствии соответствующих паттернов. Однако на практике зачастую представляют интерес правила, в которых, например, в заключении утверждается отсутствие некоторого паттерна. Такие правила принято называть *негативными* (*negative*). Если негативные ассоциативные правила сформулировать в терминах формул алгебры логики, которые могут быть поставлены в соответствие правой и левой частям правила (это можно сделать по изоморфизму алгебры множеств и алгебры логики), то в негативном ассоциативном правиле будут присутствовать формулы, содержащие литеры со знаками отрицания.

Существует много практических задач и ситуаций, где важно учитывать "негативную корреляцию" паттернов. Традиционными являются примеры из области маркетинга, когда, например, решается задача об оптимальном размещении товаров по отделам. Поскольку в этом случае неразумно помещать в одном отделе магазина товары, которые редко покупаются вместе, то их можно обнаружить путем поиска негативных ассоциативных правил на основе обработки накопленных данных о покупательских корзинах магазина.

Большое значение анализ негативных ассоциаций занимает и в других областях. Примерами являются экология (для обнаружения негативных факторов воздействия на среду), психология (здесь, например, важен анализ негативных факторов, влияющих на психологическую совместимость), анализ инвестиционных стратегий (для обнаружения негативных факторов, влияющих на перспективность инвестиционной стратегии) и многие другие классы приложений.

По существу, негативные ассоциативные правила предоставляют не менее ценное знание, чем позитивные: они описывают знания несколько иного типа по сравнению с позитивными, но которые могут быть весьма полезными.

Важная особенность негативных правил в том, что "сильные" негативные правила могут соответствовать случаям, когда формирующие их паттерны сами по себе не являются часто встречающимися. Это означает, что при поиске негативных правил нельзя пользоваться сокращением перебора, основанным на отсечении ветвей поиска, которые соответствуют редко встречающимся паттернам, а это условие составляет основное предположение всех рассмотренных ранее моделей ассоциативных правил, поскольку иначе поиск правил превращается в перебор экспоненциальной сложности. Поэтому, используя алгоритмы поиска ассоциативных правил на множестве часто встречающихся паттернов, можно потерять много очень сильных зависимости в данных.

В работе [31] формулируются следующие проблемы, которые необходимо решать при поиске негативных ассоциативных правил, как, впрочем, и при поиске сильных правил, составленных из редко встречающихся паттернов:

1). поиск редко встречающихся паттернов, которые, тем не менее, представляют интерес;

2). преодоление экспоненциального роста числа паттернов, которые необходимо анализировать при работе с редко встречающимися паттернами;

3). создание алгоритмов, которые могут одинаково эффективно обнаруживать и редко, и часто встречающиеся паттерны, представляющие интерес с позиций поиска ассоциативных правил;

4). разработка альтернативных мер для оценки полезности паттернов и ассоциативных правил.

В работе [31] рассматривается подход, в котором автор предлагает отыскивать ассоциативные правила с помощью двухступенчатой процедуры.

Сначала ассоциативные правила отыскиваются по подвыборке имеющихся данных. Её размер определяется на основании центральной предельной теоремы теории вероятности, которая позволяет оценить объем выборки в зависимости от требуемой точности оценивания некоторой статистики. Подвыборка заданного размера затем генерируется с помощью случайного механизма. Естественно, авторы [31] предполагают, что полученная таким образом подвыборка будет иметь приблизительно то же самое распределение, что и исходная. Обычно предполагается, что полученная таким образом подвыборка имеет существенно меньший объем, что позволяет быстрее отыскивать паттерны при очень слабых ограничениях на значение функции поддержки, однако оценки свойств полученных ассоциативных правил (позитивных и негативных) могут при этом получиться весьма неточными.

Затем на 2 шаге используя специальный прием, авторы сокращают количество переменных в строках базы данных, а также используют некоторые известные вероятностные соотношения, которые позволяют им быстрее решать задачу поиска негативных правил и уточнять их оценки по полной выборке. Поясним последнее несколько подробнее, поскольку именно здесь содержится основное существо предложенного подхода.

В основу сокращения перебора положено использование фактора уверенности (*certainty factor*), предложенного Е. Шортлифом (E. Shortliffe) в работе [28]:

Пусть $X$ и $Y$ – случайные события, для которых известны $p(X)$, $p(Y)$, и $p(X \cap Y)$. Тогда фактором уверенности Шортлифа называется величина, вычисляемая таким образом:

$$PR(Y/X)=$$

$$=\begin{cases}\dfrac{p(X\cap Y)-p(X)p(Y)}{p(X)\left[(1-p(Y)\right]}, \\ \text{если } p(X\cap Y)\geq p(X)p(Y) \ \textit{и} \ p(X)(1-p(Y))\neq 0; \\ \dfrac{p(X\cap Y)-p(X)p(Y)}{p(X)p(Y)}, \\ \text{если } p(X\cap Y)< p(X)p(Y) \ \textit{и} \ p(X)p(Y)\neq 0.\end{cases} \quad (12)$$

Основное свойство фактора уверенности Шортлифа выражается соотношением

$$PR(Y/X)+PR(\neg Y/X) = 0. \quad (13)$$

Понятно, что при поиске негативных ассоциативных правил вместо вероятностей, присутствующих в формулах (12), (13) используются их оценки, которые выражаются здесь в терминах функции поддержки *supp* для соответствующих событий.

С учетом свойства (13) авторы [31] формулируют основную идею их алгоритма поиска позитивных, и негативных правил для пары паттернов $X$ и $Y$ в виде следующих четырех утверждений:

1). если $supp(X \cap Y)\geq\sigma$,

$supp(X \cap Y)$ - $supp(X)$ $supp(Y) \geq \delta$

и $PR(Y/X) \geq \gamma$, то правило $X \rightarrow Y$ может быть извлечено в качестве ассоциативного правила, представляющего интерес.

2). если $supp(X \cap \neg Y)\geq\sigma$, $supp(Y)\geq\sigma$, $supp(X)\geq\sigma$,

$supp(X \cap \neg Y)$ - $supp(X)$ $supp(\neg Y)\geq\delta$

и $PR(\neg Y/X) \geq \gamma$, то правило $X \rightarrow \neg Y$ может быть извлечено в качестве ассоциативного правила, представляющего интерес.

3). если $supp(\neg X \cap Y)\geq\sigma$, $supp(Y)\geq\sigma$, $supp(X)\geq\sigma$,

$supp(\neg X \cap Y)$ - $supp(\neg X)$ $supp(Y)\geq\delta$

и $PR(Y/\neg X) \geq \gamma$, то правило $\neg X \rightarrow Y$ может быть извлечено в качестве ассоциативного правила, представляющего интерес.

4). если $supp(\neg X \cap \neg Y)\geq\sigma$, $supp(Y)\geq\sigma$, $supp(X)\geq\sigma$,

$supp(\neg X \cap \neg Y)$ - $supp(\neg X)supp(\neg Y)\geq\delta$

и $PR(\neg Y/\neg X) \geq \gamma$, то правило $\neg X \rightarrow \neg Y$ может быть извлечено в качестве ассоциативного правила, представляющего интерес.

Отметим, что три последних утверждения указывают пути извлечения негативных ассоциативных правил.

В целом этот алгоритм предлагает определенный шаг в решении задачи ассоциативных правил, однако и он обладает рядом недостатков. По сути, он распространяет модель Пятецкого-Шапиро (см. раздел 2.2) на задачу поиска негативных правил, используя ее в несколько иной форме. Различие касается использования фактора уверенности Шортлифа вместо условия (10), а это почти одно и то же. Условия 1–4, которые используются для сокращения перебора при поиске ассоциативных правил, фактически повторяют условия (10) для случаев, когда некоторые случайные события, присутствующие в посылке и/или заключении ассоциативного правила, берутся со знаком отрицания. Заслуга авторов в том, что они рассматривают вероятностную трактовку задачи поиска ассоциативных правил, что дает им возможность использовать некоторые известные свойства вероятностей для повышения эффективности алгоритмов поиска.

Иная идея предлагается в работе [27]. В ней авторы рассматривают решетку, которую можно построить на множестве элементов, формирующих паттерны, с помощью операции объединения множеств. Далее в работе предлагается использовать границу, которую в этой структуре можно построить на множестве паттернов максимальной длины с уровнем поддержки, до которого паттерны классифицируются как часто встречающиеся. Тогда все паттерны, которые лежат выше этой границы, будут относиться уже к редко встречающимся. Затем в работе предлагается алгоритм, с помощью которого следует генерировать все редко встречающиеся паттерны. Однако по своей сути этот алгоритм остается переборным с тем лишь различием, что поиск ведется выше найденной границы для часто встречающихся паттернов (однако ее построение требует поиска часто встречающихся паттернов). Авторы признаются, что метод требует хранить все такие паттерны, а это накладывает высокие требования по памяти. Другой недостаток предложенного подхода в том, что рассматриваются только редко встречающиеся *положительные* паттерны, а потому он не позволяет отыскивать негативные правила.

Хотя имеется много других публикаций на тему поиска редко встречающихся паттернов и негативных ассоциативных правил, тем не менее среди предложенных в них алгоритмов практически отсутствуют такие, которые могли бы быть полезными для практики. Большинство из них предлагает использовать различные частные

варианты негативных правил, например, диссоциативные [23], когерентные [29], исключения [22] и др. Описанный в работах [30] и [31] алгоритм, представляется наилучшим из них, однако, как уже отмечалось, он во многом наследует недостатки модели *поддержка–уверенность–зависимость*, хотя применительно к поиску негативных правил лучших алгоритмов, по видимому, до 2008 г. не предложено.

**3. Недоопределенное вероятностное пространство.** В данной работе проблема поиска ассоциативных правил, включая, а также правила причинно следственного характера рассматривается в чисто вероятностной постановке, с предположением, что оценки необходимых вероятностей с той или иной точностью могут быть получены путем обработки обучающих данных. При этом, в отличие от известных подходов, используемая модель вероятностного пространства строится иначе, чем в классическом случае, что позволяет построить единый алгоритм поиска правил для всех их вариантов. В данном разделе рассматривается модель вероятностного пространства, которая далее положена в основу решения задач, сформулированных ранее в качестве основных целей работы.

Пусть задано универсальное множество $\boldsymbol{S}$ и множество унарных операторов $B_i$, каждый из которых принимает значения "истина" в области своего задания $X_i \subseteq \boldsymbol{S}$, $i = 1, \ldots, n$, и "ложь" в противном случае, и при этом $\boldsymbol{S} = \bigcup_{i=1}^{n} X_i$ . На области задания отдельных унарных предикатов не накладывается каких либо специальных ограничений. В частности, они могут иметь (хотя и не обязательно) непустое пересечение, т. е. $X_i \cap X_j \neq \varnothing,\ X_i, X_j \subseteq \boldsymbol{S}$.

Используя в качестве системы образующих множество областей значений унарных предикатов $X_i \subseteq \boldsymbol{S}$, $i = 1, \ldots, n$, зададим булеву алгебру [1]

$$\mathcal{B}\ (\{X_i\}|_{i=1}^{n}) = <\{X_i\}|_{i=1}^{n}, \{\cup, \cap, /\}>$$

в которой $\{X_i\}|_{i=1}^{n}$–система образующих множеств, $\{\cup, \cap, /\}$ – теоретико множественные операции: операции объединения множеств, их пересечения и дополнения множества до универсального множества $\boldsymbol{S}$ соответственно.

В булевой алгебре $\mathcal{B}(\{X_i\}|_{i=1}^{n})$ существует множество подмножеств $\{A_k\}|_{k=1}^{n}$, такое, что $A_k \cap A_r = \varnothing$, $k \neq r$, $\bigcup_{k=1}^{n} A_k = \boldsymbol{S}$. Тогда в булевой алгебре $\mathcal{B}(\mathcal{A}) = <\{A_k\}|_{k=1}^{n}, \{\cup, \cap, /\}>$, которая изоморфна алгебре $\mathcal{B}(\{X_i\}|_{i=1}^{n})$, может быть определена норма [8], которая подчиняется стандартной аксиоматике:

$a_1: 0 \le |A| \le 1$; $a_2: |0| = 0$; $a_3: |1| = 1$; $a_4$: если $A_k \cap A_r = \varnothing$,

то $|A_k \cup A_r| = |A_k| + |A_r|$ для $\forall A, A_k, A_r \in \{A_k\}|_{k=1}^{n}$.

Алгебра $\mathcal{B}(\{A_k\}|_{k=1}^{n})$ с введенной нормой [8] задает вероятностное пространство

$$\mathcal{B}_{\mathcal{N}}(\mathcal{A}) = <\{A_k\}|_{k=1}^{n}, \{\cup, \cap, /\}, \mu(\{A_k\}|_{k=1}^{n})>,$$

в котором множества $\{A_k\}|_{k=1}^{n}$ играют роль элементарных случайных событий, а норма любого элемента имеет смысл его вероятности. По изоморфизму булевых алгебр $\mathcal{B}(\{A_k\}|_{k=1}^{n})$ и $\mathcal{B}(\{X_j\}|_{j=1}^{n})$ в последней также можно определить вероятностную меру $\mu(\{X_j\}|_{j=1}^{n})$ таким образом, что изоморфный образ некоторого события $A_k$ алгебры $\mathcal{B}(\{A_k\}|_{k=1}^{n})$ в алгебре $\mathcal{B}(\{X_j\}|_{j=1}^{n})$ будет иметь то же самое значение вероятности. Таким способом можно построить изоморфное (по отношению к $\mathcal{B}_{\mathcal{N}}(\mathcal{A})$) вероятностное пространство

$$\mathcal{B}_{\mathcal{N}}(\{X_j\}|_{j=1}^{n}) = <\{X_j\}|_{j=1}^{n}, \{\cup, \cap, /\}, \mu(\{X_j\}|_{j=1}^{n})>.$$

Определим в булевой алгебре $\mathcal{B}(X_j\}|_{j=1}^{n})$ множество подалгебр $\mathcal{B}^{(1)}(\{X_{i_1}\}|_{i_1=1}^{n_1}), \ldots, \mathcal{B}^{(r)}(\{X_{i_r}\}|_{i_r=1}^{n_r})$, каждая из которых имеет в качестве множества образующих некоторое подмножество системы образующих $\{X_i\}|_{i=1}^{n}$ исходной алгебры. Отметим, что при формировании подалгебр $\mathcal{B}^{(s)}(\{X_{i_s}\}|_{i_s=1}^{n_s})$, $s = 1, \ldots, r$, не

накладывается каких-либо ограничений на выбор подмножеств $\{\{X_{i_s}\}|_{i_s=1}^{n_s}\}|_{s=1}^{r}$ множества $\{X_i\}|_{i=1}^{n}$, в частности, эти подмножества могут быть пересекающимися, находиться в отношении множество–подмножество, а их объединение может не совпадать с универсальным множеством $\boldsymbol{S}$.

Напомним, что на элементах подалгебр $\mathcal{B}^{(s)}$ $(\{X_{i_s}\}|_{i_s=1}^{n_s})$, $s = 1, \ldots, r$, уже задана вероятностная мера, определенная вероятностным пространством $\mathcal{B}_{\mathcal{N}}$ $(\{X_j\}|_{j=1}^{n})$. Каждая такая подалгебра $\mathcal{B}^{(s)}$ $(\{X_{i_s}\}|_{i_s=1}^{n_s})$, $s = 1, \ldots, r$, с определенной на ней вероятностной мерой является некоторым фрагментом вероятностного пространства, задаваемого нормированной алгеброй $\mathcal{B}_{\mathcal{N}}$ $(\{X_j\}|_{j=1}^{n})$. Обозначим фрагмент вероятностного пространства, включающий в себя только элементы подалгебры $\mathcal{B}^{(s)}$ $(\{X_{i_s}\}|_{i_s=1}^{n_s})$, $s = 1, \ldots, r$, с вероятностной мерой заданной на ней символом $\mathcal{Fr}^{(s)}$, $s = 1, \ldots, r$.

Одно из важных свойств фрагмента $\mathcal{Fr}^{(s)}$, $s = 1, \ldots, r$, вероятностного пространства $\mathcal{B}_{\mathcal{N}}$ $(\{X_j\}|_{j=1}^{n})$ состоит в том, что знание вероятностей, приписанных его элементам (им соответствуют случайные события), позволяет точно вычислить вероятность любого другого случайного события, которое может быть получено с помощью теоретико–множественных операций булевой алгебры, в которых участвуют только множество элементов соответствующей подалгебры, дополненное универсальным множеством $\boldsymbol{S}$. При этом вероятности названных событий будут совпадать с вероятностями соответствующих событий, определенных в "полном" вероятностном пространстве $\mathcal{B}_{\mathcal{N}}$ $(\{X_j\}|_{j=1}^{n})$, а также и с вероятностями их изоморфных образов в вероятностном пространстве $\mathcal{B}_{\mathcal{N}}(\mathcal{A})$.

Обозначим далее символом

$$[\mathcal{B}^{(s)}, \boldsymbol{S}] = [\mathcal{B}^{(s)} (\{X_{i_s}\}|_{i_s=1}^{n_s}, \boldsymbol{S})]$$

множество элементов вероятностного пространства $\mathcal{B}_{\mathcal{N}}$ $(\{X_j\}|_{j=1}^{n})$, которые могут быть из получены из множества элементов подалгебры $\mathcal{B}^{(s)}$ $(\{X_{i_s}\}|_{i_s=1}^{n_s})$, дополненного универсальным множеством $\boldsymbol{S}$, с помощью теоретико - множественных операций $\{\cup, \cap, /\}$. [ ], как это принято [1], означают операцию замыкания множества, символ которого стоит в них, по операциям алгебры, в нашем случае – по операциям булевой алгебры. Обозначим также символом $[\mathcal{F}r^{(s)}, \boldsymbol{S}]$ фрагмент вероятностного пространства $\mathcal{B}_{\mathcal{N}}$ $(\{X_j\}|_{j=1}^{n})$, который включает в себя все элементы множества $[\mathcal{B}^{(s)}, \boldsymbol{S}]$ с вероятностной мерой, приписанной этим элементам в исходном вероятностном пространстве $\mathcal{B}_{\mathcal{N}}$ $(\{X_j\}|_{j=1}^{n})$. Таким образом, знание вероятностной меры на элементах фрагмента $\mathcal{F}r^{(s)}$ позволяет продолжить ее однозначным образом также и на элементы фрагмента $[\mathcal{F}r^{(s)}, \boldsymbol{S}]$.

Основной вывод, который можно сделать из предыдущих построений и рассуждений, состоит в том, что знание вероятностной меры на элементах фрагмента $\mathcal{F}r^{(s)}$, если только множество $[\mathcal{B}^{(s)}, \boldsymbol{S}]$ не совпадает с множеством всех событий алгебры $\mathcal{B}$ $(\{X_i\}|_{i=1}^{n})$, лишь частично определяет вероятностное пространство. Это же заключение будет иметь место и для объединения построенных множеств фрагментов, если только $\bigcup_{s=1}^{r}[\mathcal{B}^{(s)}, \boldsymbol{S}] \neq \mathcal{B}$ $(\{X_i\}|_{i=1}^{n})$. В связи с этим введем следующее определение.

*Определение 3.* Фрагмент вероятностного пространства $\mathcal{B}_{\mathcal{N}}$ $(\{X_j\}|_{j=1}^{n})$, определяемый множеством $\mathcal{F}r = \bigcup_{s=1}^{r}[\mathcal{F}r^{(s)}, \boldsymbol{S}]$ с заданной на нем вероятностной мерой будем называть *недоопределенным вероятностным пространством*. Обозначим его символом $\mathcal{F}r_{\mathcal{N}} = <\mathcal{F}r, \mu(\mathcal{F}r)>$.

Недоопределенное вероятностное пространство является абстракцией, которая имеет ясную содержательную трактовку.

Действительно, предположим, что каким-то образом, например, используя накопленные данные, можно построить оценки вероятностей некоторых событий из множества $\{X_j\}|_{j=1}^{n}$, а также некоторые из совместных вероятностей этих событий. Тогда эта информация может оказаться недостаточной для построения вероятностного пространств в том смысле, что не для всех его событий могут быть найдены вероятности.

Рассмотрим небольшой пример, поясняющий смысл понятия фрагмента вероятностного пространства и недоопределенного вероятностного пространства.

**Пример 1**. Рассмотрим две подалгебры $\mathcal{B}^{(1)}$ ($\{X_1, X_3\}$) и $\mathcal{B}^{(2)}$ ($\{X_1, X_5\}$), порождаемые подмножествами образующих $\{X_1, X_3\}$ и $\{X_1, X_5\}$ вероятностного пространства, которое полностью задается пятью образующими (базовыми событиями) и их вероятностями $p(X_i)$, $i = 1, \ldots, 5$, а также вероятностями событий, которые могут быть построены из них с помощью алгебры событий. Каждая из рассматриваемых подалгебр содержит три элемента, для которых полагаем известными (например, из экспериментальных данных) оценки вероятностной меры.

Пусть построены оценки $p(X_1)$, $p(X_3)$, $p(X_1 \cap X_3)$, $p(X_5)$ и $p(X_1 \cap X_5)$. Применяя операции булевой алгебры к множествам $\{X_1, X_3, X_1 \cap X_3, \boldsymbol{S}\}$ и $\{X_1, X_5, X_1 \cap X_5, \boldsymbol{S}\}$, можно получить следующие подмножества:

$$[X_1, X_3, \boldsymbol{S}] = \{X_1, X_3, \overline{X}_1 = \boldsymbol{S} \setminus X_1, \overline{X}_3 = \boldsymbol{S} \setminus X_3, X_1 \cap X_3, \overline{X}_1 \cap X_3, X_1 \cap \overline{X}_3, \overline{X}_1 \cap \overline{X}_3, X_1 \cup X_3, \overline{X}_1 \cup X_3, X_1 \cup \overline{X}_3, \overline{X}_1 \cup \overline{X}_3\};$$

$$[X_1, X_5, \boldsymbol{S}] = \{X_1, X_5, \overline{X}_1 = \boldsymbol{S} \setminus X_1, \overline{X}_5 = \boldsymbol{S} \setminus X_5, X_1 \cap X_5, \overline{X}_1 \cap X_5, X_1 \cap \overline{X}_5, \overline{X}_1 \cap \overline{X}_5, X_1 \cup X_5, \overline{X}_1 \cup X_5, X_1 \cup \overline{X}_5, \overline{X}_1 \cup \overline{X}_5\}.$$

Используя аксиомы нормировки и аддитивности теории вероятности для зависимых событий, можно вычислить вероятностную меру для всех остальных элементов множеств $[X_1, X_3, \boldsymbol{S}]$ и $[X_1, X_5, \boldsymbol{S}]$:

$$p(\overline{X}_1) = 1 - p(X_1),\ p(\overline{X}_3) = 1 - p(X_3),$$
$$p(\overline{X}_1 \cap X_3) = p(X_3) - p(X_1 \cap X_3),$$
$$p(X_1 \cap \overline{X}_3) = p(X_1) - p(X_1 \cap X_3),$$
$$p(\overline{X}_1 \cap \overline{X}_3) = 1 - p(X_1) - p(X_3) + p(X_1 \cap X_3),$$
$$p(X_1 \cup X_3) = p(X_1) + p(X_3) - p(X_1 \cap X_3),$$
$$p(\overline{X}_1 \cup X_3) = 1 - p(X_1) + p(X_1 \cap X_3),$$
$$p(X_1 \cup \overline{X}_3) = 1 - p(X_3) + p(X_1 \cap X_3),$$
$$p(\overline{X}_1 \cup \overline{X}_3) = 1 - p(X_1 \cap X_3);\ p(\overline{X}_5) = 1 - p(X_5),$$
$$p(\overline{X}_1 \cap X_5) = p(X_5) - p(X_1 \cap X_5),$$
$$p(X_1 \cap \overline{X}_5) = p(X_1) - p(X_1 \cap X_5),$$
$$p(\overline{X}_1 \cap \overline{X}_5) = 1 - p(X_1) - p(X_5) + p(X_1 \cap X_5),$$
$$p(X_1 \cup X_5) = p(X_1) + p(X_5) - p(X_1 \cap X_5),$$
$$p(\overline{X}_1 \cup X_5) = 1 - p(X_1) + p(X_1 \cap X_5),$$
$$p(X_1 \cup \overline{X}_5) = 1 - p(X_5) + p(X_1 \cap X_5),$$
$$p(\overline{X}_1 \cup \overline{X}_5) = 1 - p(X_1 \cap X_5).$$

Таким образом, зная вероятности пяти элементов из множества $[X_1, X_3, \boldsymbol{S}] \cup [X_1, X_5, \boldsymbol{S}]$, можно вычислить вероятности еще 17 элементов вероятностного пространства, задаваемого нормированной булевой алгеброй $\boldsymbol{\mathcal{B}}_{\mathcal{N}}(\{X_j\}|_{j=1}^{5})$. Пользуясь вероятностями уже известных событий, для остальных событий вероятностного пространства, задаваемого нормированной булевой алгеброй $\boldsymbol{\mathcal{B}}_{\mathcal{N}}(\{X_j\}|_{j=1}^{5})$, можно получить интервальные оценки [4].

Обратим внимание на различия и особенности введенного вероятностного пространства по сравнению с классическим вероятностным пространством, которое вводится с помощью

пространства элементарных событий с классическим аксиомами нормировки, аддитивности (и аксиомы условной вероятности):

1) базовые события, например, $X_i \subseteq \boldsymbol{S}$ и другие множества, принадлежащие алгебре $\mathcal{B}$ ($\{X_i\}|_{i=1}^{n}$), могут быть зависимыми, т. е. базовые события введенного вероятностного пространства не являются элементарными;

2) классическая модель вероятностного пространства является частным случаем вероятностного пространства, задаваемого нормированной булевой алгеброй $\mathcal{B}_{\mathcal{N}}$ ($\{X_j\}|_{i=1}^{n}\}$), когда $\{X_i\}_{i=1}^{n}$ является множеством элементарных событий;

3) для вероятностного пространства $\mathcal{B}_{\mathcal{N}}$ ($\{X_j\}|_{i=1}^{n}\}$) может быть построена изоморфная модель классического вероятностного пространства.

Заметим, что в основу построения введенного здесь вероятностного пространства положена модель, описанная в работах [15, 16, 18].

Одним из достоинств модели вероятностного пространства, введенной в работе [15], является возможность корректного перехода от нее к модели вероятностной логики

$$\mathcal{L}_{\mathcal{N}} = <\{\mathfrak{I}, \mu(\mathfrak{I})\}>,$$

где $\mathfrak{I}$ –алгебра формул,

$$\mathfrak{I} = <\{B_i\}_{i=1}^{n}, \{\vee, \wedge, \neg\}>,$$

где $\vee, \wedge, \neg$ – логические связки: дизъюнкция, конъюнкция и отрицание соответственно, а $\mu(\mathfrak{I})$ – множество вероятностных мер, заданных на элементах алгебры формул.

Изоморфизм $\varphi$: $\mathcal{B}_{\mathcal{N}} \Leftrightarrow \mathcal{L}_{\mathcal{N}}$ соответствует взаимно однозначному отображению $\varphi$ алгебры множеств $<\{X_j\}|_{j=1}^{n}\}$, $\{\cup, \cap, /\}>$ на алгебру формул $\mathfrak{I}$, в котором

$$\varphi: X_i \leftrightarrow B_i|_{j=1}^{n}, \cup \leftrightarrow \vee, \cap \leftrightarrow \wedge, \setminus \leftrightarrow \neg,$$

а также взаимно однозначному отображению мер, т. е.

$$\varphi: [\forall\, Y_j \in \mathcal{B}\ (\{X_i\}|_{j=1}^{n}),\ \forall\, F_j \in \mathfrak{I},\ F_j = \varphi(Y_j),\ A_j = \varphi^{-1}(F_j)]: \mu(F_j) = \mu(Y_j)]. \quad (14)$$

Заметим, что в качестве изоморфных образов системы образующих $\{X_j\}|_{j=1}^{n}\}$, задающих вероятностное пространство $\mathcal{B}_{\mathcal{N}}$, выступают области истинности унарных предикатов $B_i|_{j=1}^{n}$, а множество этих унарных предикатов является системой образующих в пространстве $\mathcal{L}_{\mathcal{N}}$. Обратим внимание также на то, что в формуле (13) произвольный элемент подмножества конечной алгебры множеств (событий) $\mathcal{B}$ ($\{X_i\}|_{j=1}^{n}$) обозначен символом $Y_j$ вместо символа $X_i$ для того, чтобы подчеркнуть, что речь идет о произвольном элементе алгебры событий $\mathcal{B}$ ($\{X_i\}|_{j=1}^{n}$), а не обязательно только о базовых событиях $X_i$, из которых с помощью операций множества $\{\cup, \cap, \setminus\}$ формируется все семейство множеств алгебры $\mathcal{B}$ ($\{X_i\}|_{j=1}^{n}$). Основные свойства этой логики описаны в работе [15].

Очевидно, что частью этого изоморфизма является изоморфизм недоопределенного вероятностного пространства на соответствующий фрагмент вероятностной логики $\mathcal{L}_{\mathcal{N}} = <\{\mathfrak{I}, \mu(\mathfrak{I})\}>$. Пусть $\mathbf{\Phi} = \bigcup_{s=1}^{r}[\mathbf{\Phi}^{(s)}, \boldsymbol{S}]$ есть изоморфный образ множества $\bigcup_{s=1}^{r}[\mathcal{F}\boldsymbol{r}^{(s)}, \boldsymbol{S}]$ в алгебре формул $\mathfrak{I}$, т. е.

$$\varphi(\bigcup_{s=1}^{r}[\mathcal{F}\boldsymbol{r}^{(s)}, \boldsymbol{S}]) = \bigcup_{s=1}^{r}[\mathbf{\Phi}^{(s)}, \boldsymbol{S}].$$

Тогда аналогично недоопределенному вероятностному пространству, введенному в определении 1, можно ввести понятие недоопределенной вероятностной логики:

*Определение 4*. Назовем фрагмент $\mathbf{\Phi}_{\mathcal{N}}$ вероятностной логики $\mathcal{L} = <\{\mathfrak{I}, \mu(\mathfrak{I})\}>$, заданный на множестве формул $\mathbf{\Phi} = \bigcup_{s=1}^{r}[\mathbf{\Phi}^{(s)}, \boldsymbol{S}]$, недоопределенной вероятностной логикой.

Сделаем важное замечание о том, что полезное можно получить, вводя изоморфизм φ. В вероятностной логике часто бывает необходимость вычислять вероятности некоторых формул, когда заданы вероятности других формул. Это выполнить гораздо проще и естественнее, если соответствующие вычисления сделать для

изоморфных элементов вероятностного пространства $\mathcal{B}_{\mathcal{N}}$ , пользуясь стандартной вероятностной аксиоматикой.

Однако использование булевой алгебры для задания вероятностного пространства приводит к избыточности и его можно задать более экономным образом. Покажем это.

Из теории структур [1] известно, что любая булева алгебра, а значит и булева алгебра $\mathcal{B}$ $(\{X_i\}|_{i=1}^{n})$, является решеткой, в которой можно естественным образом ввести отношение порядка на множестве его элементов. В алгебре $\mathcal{B}$ $(\{X_i\}|_{i=1}^{n})$ в качестве такого отношения выступает отношение включения семейства множеств, которые образуют ее основание. Обозначим решетку множеств, частично упорядоченных таким образом, символом $\mathcal{R}$ ,

$$\mathcal{R} = <\Re, \leq>,$$

где $\Re$ – множество все подмножеств булевой алгебры $\mathcal{B}$ $(\{X_i\}|_{i=1}^{n})$, а порядок "$\leq$" определен через отношение включения множеств "$\subseteq$".

Каждому элементу этой решетки, порожденной булевой алгеброй $\mathcal{B}$ $(\{X_i\}|_{i=1}^{n})$, можно поставить в соответствие вероятностную меру, определенную в нормированной булевой алгеброй $\mathcal{B}_{\mathcal{N}}$ $(\{X_j\}|_{i=1}^{n}\})$, которую далее будем называть нормированной решеткой. Обозначим ее символом $\mathcal{R}_{\mathcal{N}} = <\Re, \leq, \mu(\Re)>$.

Используя нормированную решетку $\mathcal{R}_{\mathcal{N}} = <\Re, \leq, \mu(\Re)>$, можно построить две полурешетки относительно каждой из операций булевой алгебры, операции пересечения множеств и их объединения, примененных к множеству образующих $\{X_i\}|_{i=1}^{n}$. Обозначим первую из них символом $\mathcal{R}_{\mathcal{N}}^{\cap}$ (ее принято также называть нижней полурешеткой), а вторую – символом $\mathcal{R}_{\mathcal{N}}^{\cup}$ (верхняя полурешетка). Оказывается, что каждая из них полностью задает вероятностное пространство. В частности, этот факт использован в примере 1, когда вероятности объединения событий вычисляются через вероятности базовых событий и вероятности их пересечений. Верхняя и нижняя полурешетки, как показано в следующем разделе, служат удобными

структурами для представления дискретного распределения. При этом понятие фрагмента вероятностного пространства, введенное в данном разделе, соответствуют в структуре $\mathcal{R}_{\mathcal{N}}$ понятию подрешетки. Ее компоненты в полурешетках $\mathcal{R}_{\mathcal{N}}^{\cap}$ и $\mathcal{R}_{\mathcal{N}}^{\cup}$ задают те же самые недоопределенные вероятностные пространства, которые введены в определениях 3 и 4.

Аналогичные утверждения справедливы относительно вероятностной логики $\mathcal{L}_{\mathcal{N}} = <\{ \mathfrak{I},\ \mu(\mathfrak{I})\ \}>$ и ее фрагмента $\mathbf{\Phi}_{\mathcal{N}}$.

Заметим, что фрагменты $\bigcup_{s=1}^{r}[\mathcal{F}r^{(s)}, \boldsymbol{S}]$ и $\bigcup_{s=1}^{r}[\mathbf{\Phi}^{(s)}, \boldsymbol{S}]$ уже не являются решетками; они в общем случае только частично упорядоченные множества.

**4. Ассоциативная Байесовская сеть.** Понятие АБС как формы представления недоопределенного вероятностного пространства, заданного на конечном множестве случайных событий, в менее строгой форме впервые введено в работах [2, 3, 17]. Название "алгебраическая байесовская сеть", выбранное в этих работах для введенной модели вероятностного пространства, имело цель подчеркнуть алгебраическое происхождение этой модели. В данной работе для АБС используется несколько иное название для того, чтобы подчеркнуть акцент данной работы на вопросах анализа ассоциаций. Однако здесь эта модель рассматривается в более общей форме. Введем формально понятие АБС.

*Определение 5.* Изоморфный образ недоопределенного вероятностного пространства $\mathcal{F}r_{\mathcal{N}} = <\mathcal{F}r, \mu(\mathcal{F}r)>$ (а также $\mathbf{\Phi}_{\mathcal{N}} = <\mathbf{\Phi}, \mu(\mathbf{\Phi})>$ в нижнюю полурешетку $\mathcal{R}_{\mathcal{N}}^{\cap}$ будем называть ассоциативной (алгебраической) байесовской сетью.

Аналогично можно определить *двойственную* байесовскую сеть как изоморфный образ указанных в определении 5 вероятностных моделей в верхнюю полурешетку. Далее в данной работе, главным образом будет использоваться АБС, введенная в определении 5, хотя в ряде прикладных задач совместное использование обеих моделей АБС достаточно плодотворно. Примером такого приложения является анализ зависимости и меры разнообразия правил в задачах объединения решений [7].

**Пример 2.** Рассмотрим пример АБС (рис. 3), представленной в виде диаграммы Хассе с вероятностями, приписанными ее узлам. В этой модели имеется шесть базовых элементов $X_1, \ldots, X_6$, на которых построены четыре фрагмента, заданных множествами образующих $\{X_1, X_2, X_3\}$, $\{X_2, X_3, X_4\}$, $\{X_4, X_5\}$ и $\{X_5, X_6\}$.

Заметим, что здесь не обсуждается вопрос о том, каким образом эта АБС построена и почему она имеет именно такой вид. В дальнейшем при рассмотрении вопросов, связанных с практическим использованием модели АБС эти аспекты пояснены. Отметим, что АБС в терминах фрагмента вероятностной логики, изоморфная данной АБС по отображению $\varphi$, будет

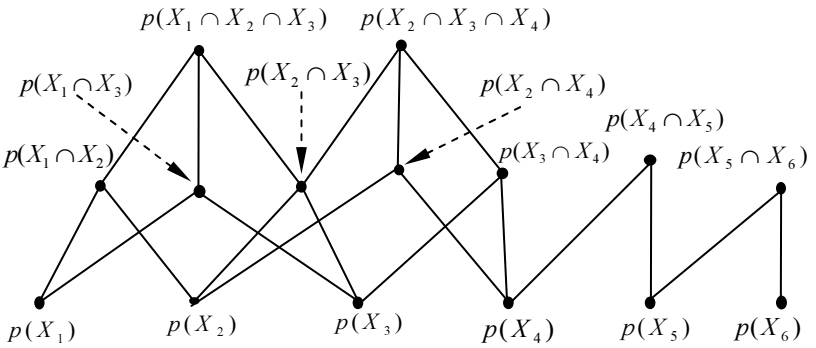


Рис. 3. Пример Ассоциативной байесовской сети.

выглядеть точно так же, однако вместо символов множеств $X_1, \ldots, X_6$ базовые узлы АБС должны быть помечены символами унарных предикатов $B_i$, $i = 1,\ldots,5$, областями истинности которых являются множества $X_1, \ldots, X_6$, а вместо теоретико–множественной операции $\cap$ нужно было бы соединять предикаты $B_i$, $i = 1,\ldots,5$ символом конъюнкции $\&$.

Будем называть далее множество образующих отдельного фрагмента его *основанием*, а его мощность – *порядком фрагмента*.

Представление АБС в виде диаграммы Хассе позволяет наглядно представить множество событий, для которых имеются оценки вероятностей. Если некоторый более высокий узел АБС соответствует пересечению *независимых* событий, то его можно не представлять в ней.

Заметим, что один из возможных вариантов продолжения АБС может быть основан на гипотезе независимости *минимальных* узлов сети или на гипотезе их условной независимости.

Другое достоинство АБС состоит в том, что она позволяет достаточно компактно и в структурированной форме представить имеющуюся вероятностную информацию. Действительно, на основании вероятностной аксиоматики она позволяет вычислять вероятности всех подмножеств (событий), которые могут быть построены из событий каждого отдельного фрагмента с помощью операций алгебры событий. Покажем это для фрагментов второго и третьего порядков, представленных на рис. 4 [4], используя обозначения, описанные далее.

*Вероятностная аксиоматика для фрагмента второго порядка.* Пусть $X_i$ и $X_j$ – два базовых элемента АБС, формирующие в ней максимальный узел второго порядка, для которого известны вероятности $p(X_i)$, $p(X_j)$ и $p(X_iX_j)$[1]. Тогда справедливы следующие соотношения (по умолчанию полагаем, что все вероятности неотрицательны и принадлежат интервалу [0, 1], не вводя соответствующие аксиомы):

*Аксиомы нормировки:*

$$p(\overline{X}_i) = 1 - p(X_i),$$

$$p(\overline{X}_i\overline{X}_j) = 1 - p(X_i) - p(X_j) + p(X_iX_j);$$

*Аксиомы аддитивности:*

$$p(\overline{X}_iX_j) = p(X_j) - p(X_iX_j),$$

$$p(X_i\overline{X}_j) = p(X_i) - p(X_iX_j);$$

*Формулы включения–исключения:*

$$p(X_i \cup X_j) = p(X_i) + p(X_j) - p(X_iX_j),$$

$$p(\overline{X}_i \cup X_j) = 1 - p(X_i) + p(X_iX_j),$$

$$p(X_i \cup \overline{X}_j) = 1 - p(X_j) + p(X_iX_j),$$

$$p(\overline{X}_i \cup \overline{X}_j) = 1 - p(X_iX_j);$$

[1] Далее для краткости записи символ пересечения множеств, а также символ конъюнкции между унарными предикатами (в вероятностной логике) будет опускаться.

Вероятности любых других сложных событий (формул), содержащих не более двух базовых событий, могут быть вычислены через вероятности событий, входящих в соответствующий фрагмент АБС второго порядка.

*Вероятностная аксиоматика для фрагмента АБС третьего порядка.* В нем АБС задает следующие вероятности: $p(X_i)$, $p(X_j)$, $p(X_k)$, $p(X_iX_j)$, $p(X_iX_k)$, $p(X_jX_k)$ и $p(X_iX_jX_k)$. Для вероятностей любых событий, которые содержат не более двух базовых элементов, справедлива аксиоматика, приведенная для фрагмента второго порядка с соответствующей подстановкой индексов событий.

Дополнительная аксиоматика для вероятностей событий, содержащих не менее трех базовых элементов, включает в себя следующие соотношения:

*Аксиома нормировки:*

$$p(\overline{X}_i\overline{X}_j\overline{X}_k) = 1 - p(X_i) - p(X_j) - p(X_k) + p(X_iX_j) + p(X_iX_k) + p(X_jX_k) - p(X_iX_jX_k);$$

*Аксиомы аддитивности:*

$$p(\overline{X}_iX_jX_k) = p(X_jX_k) - p(X_iX_jX_k),$$

$$p(X_i\overline{X}_jX_k) = p(x_ix_k) - p(X_iX_jX_k),$$

$$p(X_iX_j\overline{X}_k) = p(X_iX_j) - p(X_iX_jX_k),$$

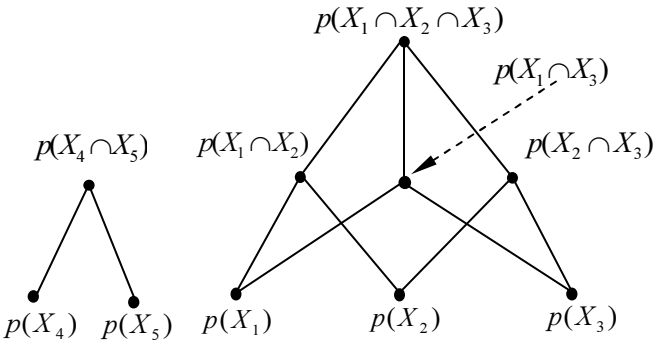


Рис.4. Примеры полных фрагмента АБС второго (слева) и третьего (справа) порядков

$$p(\bar{X}_i \bar{X}_j X_k) = p(X_k) - p(X_i X_k) - p(X_j X_k) + p(X_i X_j X_k),$$
$$p(\bar{X}_i X_j \bar{X}_k) = p(X_j) - p(X_i X_j) - p(X_j X_k) + p(X_i X_j X_k),$$
$$p(X_i \bar{X}_j \bar{X}_k) = p(X_i) - p(X_i X_j) - p(X_i X_k) + p(X_i X_j X_k);$$

*Формулы включения–исключения:*

$$p(X_i \cup X_j \cup X_k) = p(X_i) + p(X_j) + p(X_k) - p(X_i X_j) -$$
$$p(X_i X_k) - p(X_j X_k) + p(X_i X_j X_k),$$
$$p(\bar{X}_i \cup X_j \cup X_k) = 1 - p(X_i) + p(X_i X_j) + p(X_i X_k) - p(X_i X_j X_k),$$
$$p(X_i \cup \bar{X}_j \cup X_k) = 1 - p(X_j) + p(X_i X_j) + p(X_j X_k) - p(X_i X_j X_k),$$
$$p(X_i \cup X_j \cup \bar{X}_k) = 1 - p(X_k) + p(X_i X_k) + p(X_j X_k) - p(X_i X_j X_k),$$
$$p(\bar{X}_i \cup \bar{X}_j \cup X_k) = 1 - p(X_i X_j) + p(X_i X_j X_k),$$
$$p(\bar{X}_i \cup X_j \cup \bar{X}_k) = 1 - p(X_i X_k) + p(X_i X_j X_k),$$
$$p(X_i \cup \bar{X}_j \cup \bar{X}_k) = 1 - p(X_j X_k) + p(X_i X_j X_k),$$
$$p(\bar{X}_i \cup \bar{X}_j \cup \bar{X}_k) = 1 - p(X_i X_j X_k).$$

Вероятности любых других сложных событий (формул), содержащих не более трех базовых событий, могут быть также вычислены через вероятности событий, входящих в соответствующий фрагмент сети третьего порядка

Аналогичным образом записываются аксиомы для узла АБС четвертого порядка. Они могут быть найдены на веб–сайте [32].

Резюмируя содержание данного раздела, следует подчеркнуть, что АБС является удобным и притом весьма компактным представлением вероятностной информации о недоопределенном вероятностном пространстве, которое задается множеством фрагментов различных уровней. Далее показано, что такое представление позволяет решать ряд задач анализа ассоциаций в интересах практически важных приложений из области интеллектуального анализа данных, причем среди них имеются и такие, которые традиционно рассматриваются как проблемные, и для которых пока не найдено эффективных моделей и алгоритмов решения.

**5. Ассоциативный и причинный анализ на основе модели АБС.**

**5.1. Постановка задачи.** В данной работе для решения задач анализа ассоциаций используется вероятностная трактовка постановок задач, решаемых в ассоциативном анализе, и для их решения привлекается неклассическая модель вероятностного пространства, в частности, модель недоопределенного вероятностного пространства.

Целью данного раздела является краткое описание основных идей, позволяющих эффективно использовать модель АБС для построения ассоциативных правил. Новые черты постановки задачи и алгоритма решения состоят в том, что она дает возможность:

1) отказаться от предположения о том, что ассоциативные правила отыскиваются только на множестве часто встречающихся паттернов. Это позволяет использовать один и тот же алгоритм для поиска и часто, и редко встречающихся ассоциативных правил, которые отражают достаточно сильные ассоциативные зависимости. Иначе говоря, в данной работе рассматривается модель *уверенность–зависимость–причинность* вместо традиционных моделей типа *поддержка–уверенность* или *поддержка–уверенность–зависимость* (см. разделы 2.1 и 2.2);

2) отыскивать негативные правила без дополнительных проходов по базе данных, из которых извлекаются ассоциативные правила. Модель АБС позволяет это делать благодаря использованию стандартной вероятностной аксиоматики, которая может вызываться для каждого узла АБС. Примеры такой аксиоматики для фрагментов АБС второго и третьего порядка были приведены в конце раздела 4;

3) отказаться от перебора вариантов ассоциативных правил, которые могут быть построены для найденного часто встречающегося паттерна. Это достигается благодаря использованию наряду с мерой уверенности новой метрики, которая известна в теории вероятности как *коэффициент регрессии случайных событий*. Эта метрика позволяет оценивать непосредственно меру причинно – следственной зависимости паттернов.

Рассмотрим постановку задачи поиска ассоциаций в базе данных. Будем полагать, как и в разделе 2, что база данных содержит множество записей (примеров). Они могут быть представлены плоской таблицей, в которой каждому столбцу поставлено в соответствие имя одного из узлов АБС. База данных может быть также транзакционной.

В начальном состоянии АБС состоит только из однолитерных (базовых) узлов, и каждому такому узлу поставлено в соответствие либо имя множества (случайного события) $X_i \subseteq \boldsymbol{S}$, $i \in \{1, \ldots, n\}$,

либо имя унарного предиката $B_i$, $i \in \{1, \ldots, n\}$. В первой постановке задачи в каждом примере базы данных фиксируется появление или отсутствие события $X_i$, а во второй соответствующая позиция будет фиксировать истинностное значение унарного предиката $B_i$, $i \in \{1, \ldots, n\}$. Задача состоит в том, чтобы в процессе работы алгоритма построить АБС, узлы которой являются паттернами (не обязательно часто встречающимися), которые используются для генерации ассоциативных правил, удовлетворяющих условиям, которые перечислены в постановке задачи, рассматриваемой далее.

*Постановка задачи (модель уверенность–зависимость–причинность).* Пусть $X = X_{i_1} \cap X_{i_2} \cap \ldots \cap X_{i_k}$ и $Y = X_{j_1} \cap X_{j_2} \cap \ldots \cap X_{j_s}$ – два паттерна, причем $X$ и $Y$ не содержат общих узлов, т. е. среди индексов $i_*$ и $j_*$ нет одинаковых. Ассоциативным правилом называется отношение $X \to Y$, заданное на паре паттернов $X$ и $Y$, для которых выполнены следующие условия:

1. $p(Y/X) \geq \gamma_{\min}$;
2. $\left| \dfrac{P(X \cap Y)}{P(X)P(Y)} - 1 \right| \geq \delta_{\min}$.
3. $R(X,Y) = p(Y/X) - p(Y/\overline{X}) \geq \beta_{\min}$.

В приведенной постановке задачи условия 1 и 2 трактуются точно так же, как в постановке задачи типа *поддержка–уверенность–зависимость*, приведенной в разделе 2.2. Условие 3, однако, нуждается в дополнительных пояснениях.

Коэффициент регрессии случайного события $X$ на случайное событие $Y$ вычисляется как разность условных вероятностей события $Y$ при наступлении и при отсутствии события $X$. После несложных преобразований для величины $R(X,Y)$ можно получить следующее выражение:

$$R(X,Y) = \frac{p(X \cap Y) - p(x)p(Y)}{p(X)[1 - p(x)]},$$

так что значение регрессии вычисляется через вероятности событий, которые могут быть найдены, если в АБС имеются соответствующие

узлы. Запишем также выражение для условной вероятности, которая присутствует в условии пункта 1:

$$p(Y / X) = \frac{p(X \cap Y)}{p(x)}.$$

Дадим теперь краткое описание алгоритма поиска ассоциативных правил в модели *уверенность–зависимость–причинность*, сформулированной ранее в постановке задачи. В следующем разделе продемонстрируем этот алгоритм на содержательном примере, в котором найденные ассоциативные правила используются для решения задачи предсказания.

**5.2. Алгоритм поиска ассоциативных и причинных правил в модели *уверенность–зависимость–причинность*.** Предполагается, что в качестве узлов АБС выступают имена переменных $X_i \in \boldsymbol{S}$, $i \in \{1, ..., n\}$. Для каждой такой переменной имеется процедура, которая определяет ее значение (произошло событие или не произошло), и в записях базы данных (в примерах) уже записаны значения этих переменных. В примере, приведенном в следующем разделе, поясняется, каким образом могут строиться множества $X_i \in \boldsymbol{S}$, $i \in \{1,...,n\}$ и соответствующие им унарные предикаты. Алгоритм может строиться в терминах подмножеств $X_i$, а также в терминах унарных предикатов $B_i$. Он состоит из следующих этапов.

1. *Поиск оценок вероятностей узлов* $X_i \in \boldsymbol{S}$, $i \in \{1, \dots, n\}$. Для этого используется база данных. Пусть $p(X_i)$ – найденные оценки вероятностей узлов.

2. *Алгоритм поиска и фильтрации паттернов второго порядка.*

2.1. Если все пары $< X_i, X_{\mathrm{k}} >$ рассмотрены, то переход на п.3.

2.2. Для очередной пары $< X_i, X_{\mathrm{k}} >$, $i \in \{1, \dots, n\}$, $i > k$, по данным обучающей выборки находятся поочередно значения эмпирических вероятностей их совместного проявления $p(X_i X_{\mathrm{k}})$.

2.3. *Фильтрация 1*. Оценить значения критерия независимости (проверка выполнимости порогового условия 2 в постановке задачи):

$$I(X_i, X_{\mathrm{k}}) = | p(X_i X_{\mathrm{k}}) - p(X_i)\ p(X_k) | / \qquad (15)$$

$$/[\, p(X_i)\ p(X_k)\,] \geq \delta_{\min};$$

$$I(\overline{X}_i, X_{\mathrm{k}}) = |\, p(\overline{X}_i X_{\mathrm{k}}) - p(\overline{X}_i)\ p(X_k)\,|/ \qquad (16)$$

$$/[\, p(\overline{X}_i)\ p(X_k)\,] \geq \delta_{\min};$$

$$I(X_i, \overline{X}_{\mathrm{k}}) = |\, p(X_i \overline{X}_{\mathrm{k}}) - p(X_i)\ p(\overline{X}_k)\,|/ \qquad (17)$$

$$/[\, p(X_i)\ p(\overline{X}_k)\,] \geq \delta_{\min};$$

$$I(\overline{X}_i, \overline{X}_{\mathrm{k}}) = |\, p(\overline{X}_i \overline{X}_{\mathrm{k}}) - p(\overline{X}_i)\ p(\overline{X}_k)\,|/ \qquad (18)$$

$$[\, p(\overline{X}_i)\ p(\overline{X}_k)\,] \geq \delta_{\min}.$$

Все вероятности, входящие в формулы (15)–(18) могут быть вычислены через три вероятности: $p(X_i)$, $p(X_k)$ и $p(X_i X_{\mathrm{k}})$, которые вычислены в п.1 и 2.1.Заметим, что в литературе по теории вероятностей величина $I(X_i, X_{\mathrm{k}})$ называется *коэффициентом корреляции случайных событий* (в отличие от обычно используемой величины – *коэффициента корреляции случайных величин)* которая изменяется в пределах [ – 1, + 1]. Близость модуля ее значения к нулю является признаком независимости соответствующих случайных событий.

Если неравенства, заданные формулами (15) – (18) выполнены, то пара $< X_i\ X_{\mathrm{k}}, p(X_i X_{\mathrm{k}}) >$ является кандидатом на включение в АБС. Иначе она из дальнейшего анализа исключается.

2.4. *Фильтрация 2*. Вычислить значение функции уверенности и проверить пороговое условие 1 в постановке задачи:

$$p\,(X_i / X_k) = p(X_i X_{\mathrm{k}}) / p(X_k) \geq \gamma_{\min}\ ,$$

$$p\,(X_i / \overline{X}_k) = p(X_i \overline{X}_{\mathrm{k}}) / p(\overline{X}_k) \geq \gamma_{\min}\ ,$$

$$p\,(\overline{X}_i / X_k) = p(\overline{X}_i X_{\mathrm{k}}) / p(X_k) \geq \gamma_{\min}\ ,$$

$$p\,(\overline{X}_i / \overline{X}_k) = p(\overline{X}_i \overline{X}_{\mathrm{k}}) / p(\overline{X}_k) \geq \gamma_{\min}\ ,$$

$$p\,(X_k / X_i) = p(X_i X_{\mathrm{k}}) / p(X_i) \geq \gamma_{\min}\ ,$$

$$p\,(X_k / \overline{X}_i) = p(\overline{X}_i X_{\mathrm{k}}) / p(\overline{X}_i) \geq \gamma_{\min}\ ,$$

$$p\,(\overline{X}_k / X_i) = p(X_i \overline{X}_{\mathrm{k}}) / p(X_i) \geq \gamma_{\min}\ ,$$

$$p\,(\overline{X}_k / \overline{X}_i) = p(\overline{X}_i \overline{X}_{\mathrm{k}}) / p(\overline{X}_i) \geq \gamma_{\min}\ .$$

Если хотя бы одно из этих условий выполняется, то переход на п. 2.5. Иначе пара $< X_i\ X_{\text{k}}, p(X_i X_{\text{k}}) >$ из дальнейшего анализа исключается и выполняется переход на п. 2.1.

2.5. *Фильтрация 3*. Вычисляются значения коэффициентов регрессии:

$$R(X_i, X_k) = p(X_i X_{\text{k}}) / p(X_k) - p(X_i \overline{X}_{\text{k}}) / p(\overline{X}_k) =$$
$$= [\, p(X_i X_{\text{k}}) - p(X_i)\ p(X_k)\,] / [\, p(X_i)\,(1 - p(X_i))\,],$$
$$R(\overline{X}_i, X_k) = - R(X_i, X_k),$$
$$R(X_i, \overline{X}_k) = - R(X_i, X_k),$$
$$R(\overline{X}_i, \overline{X}_k) = R(X_i, X_k),$$

(т. е. нужно вычислить только одно число и сохранить одно значение).

Аналогично рассчитываются

$$R(X_k, X_i) = [\, p(X_i X_{\text{k}}) - p(X_i)\ p(X_k)\,] / [\, p(X_k)\,(1 - p(X_k))\,],$$
$$R(\overline{X}_k, X_i) = - R(X_k, X_i),$$
$$R(X_k, \overline{X}_i) = - R(X_k, X_i),$$
$$R(X_k, \overline{X}_i) = R(X_k, X_i).$$

Если $|R(X_i, X_k| \geq \delta_{\min}$ и $|R(X_i, X_k| > |R(X_k, X_i)|$, то сохраняется ассоциативное правило $X_i \to X_k$. Если $|R(X_k, X_i) \geq \delta_{min}$ и $|R(X_k, X_i)| > |R(X_i, X_k|$, то сохраняется ассоциативное правило $X_k \to X_i$. Иначе узлу $X_i X_{\text{k}}$ не ставится в соответствие ассоциативное правило и этот узел далее в АБС не рассматривается.

Переход на п.2.1.

*Примечание*. В результате работы алгоритма на втором шаге в АБС, дополнительно к узлам первого порядка будет сгенерировано некоторое число узлов второго порядка, каждому из которых будет уже поставлено в соответствие ассоциативное правило, которое на последующих шагах участвует в построении ассоциативных правил с более длинными посылками.

3. *Алгоритм поиска и фильтрации паттернов третьего порядка.* Этот алгоритм по своей сути аналогичен алгоритму

построения и фильтрации узлов второго порядка. Основное различие состоит в том, что при этом используются другие формулы для вычисления критериев отбора (фильтрации) узлов АБС.

Перед началом фильтрации работает алгоритм генерации трехлитерных узлов, которые являются кандидатами на включение в АБС и порождение ассоциативных правил. Этот алгоритм выбирает очередной узел из множества двухлитерных узлов, оставшихся в списке после фильтрации на шаге 2, и поочередно добавляет к нему по одному из однолитерных узлов, которых нет в выбранной паре. Далее полученный список фильтруется, как описывается далее [1].

3.1. Если все тройки $< X_i \; X_j \; X_{\mathrm{k}} >$ рассмотрены, то переход на п.4.

3.2. Для очередной тройки $< X_i \; X_j \; X_{\mathrm{k}} >$, $i \in 1, \ldots, n$, $i > j$, $j > k$, в которой пара $X_i \; X_j$ выбирается из числе тех, что отобраны на шаге 2, а третья – из множества оставшихся литер, по данным обучающей выборки находится значение эмпирической вероятности их совместного проявления в базе данных $p(X_i X_j X_{\mathrm{k}})$. Все остальные вероятности, которые будут нужны для дальнейших расчетов, уже получены на шаге 2.

3.3. *Фильтрация 1*. Оценить значения критерия независимости (проверка выполнимости порогового условия 2 в постановке задачи):

$$I(X_i \; X_j, X_{\mathrm{k}}) = | \, p(X_i X_i X_{\mathrm{k}}) \; - \; p(X_i X_j) \; p(X_k) \, | / \\ / [\, p(X_i X_j) \; p(X_k) \,] \geq \delta_{\min}, \tag{19}$$

$$I(\neg (X_i \; X_j), X_{\mathrm{k}}) = | \, p( \, \neg (X_i \; X_j) X_k \,) - \\ - \; p( \, \neg (X_i \; X_j) \, p(X_k) \, | / [\, p( \, \neg (X_i \; X_j) \, p(X_k) \,] \geq \delta_{\min}, \tag{20}$$

$$I(X_i \; X_j, \overline{X}_{\mathrm{k}}) = | \, p(X_i X_{\mathrm{j}} \overline{X}_k) \; - \; p \; (X_i \; X_j) \, p(\overline{X}_k) \, | / \\ - [\, p(X_i X_j) \; p(\overline{X}_k) \,] \geq \delta_{\min}, \tag{21}$$

[1] Естественно, что программная реализация процедуры генерации трехлитерных узлов–потенциальных кандидатов для поиска на их основе ассоциативных правил может быть построена более экономно. Но здесь описывается скорее основная идея поиска ассоциаций, чем рабочий алгоритм, а потому вопросы экономии памяти и/или процессорного времени оставлены в стороне.

$$I(\neg(X_i X_j), \overline{X}_{\mathrm{k}}) = |\, p(\neg(X_i X_j)\overline{X}_k) - p(\neg(X_i X_j)\, p(\overline{X}_k)|\,/[\, p(\neg(X_i X_j)\, p(\overline{X}_k)] \geq \delta_{\min}. \quad (22)$$

Часть вероятностей, входящих в формулы (19) – (22) могут быть вычислены через пять вероятностей: $p(X_i)$, $p(X_j)$, $p(X_k)$ $p(X_i X_{\mathrm{j}})$ и $p(X_i X_i X_{\mathrm{k}})$, найденные ранее в п. 1, 2.2 и 3.2. Для вычисления других вероятностей, входящих в формулы (19)–(22), можно воспользоваться аксиоматикой трехлитерного узла АБС, приведенной в разделе 4.

Если неравенства (19)–(22) выполнены, то тройка $< X_i\, X_j\, X_{\mathrm{k}}, p(X_i X_j X_{\mathrm{k}}) >$ является кандидатом на включение в АБС в качестве узла третьего порядка. Иначе она из дальнейшего анализа исключается.

3.4. *Фильтрация 2*. Вычислить значение функции уверенности и проверить пороговое условие 1 в постановке задачи:

$$p(X_i\, X_j / X_k) = p(X_i X_j X_{\mathrm{k}}) / p(X_k) \geq \gamma_{\min},$$
$$p(X_i\, X_j / \overline{X}_k) = p(X_i X_j \overline{X}_{\mathrm{k}}) / p(\overline{X}_k) \geq \gamma_{\min},$$
$$p(\neg(X_i\, X_j) / X_k) = p(\neg(X_i X_j) X_{\mathrm{k}}) / p(X_k) \geq \gamma_{\min},$$
$$p(\neg(X_i\, X_j) / \overline{X}_k) = p(\neg(X_i X_j) \overline{X}_{\mathrm{k}}) / p(\overline{X}_k) \geq \gamma_{\min},$$
$$p(X_k / X_i\, X_j) = p(X_i X_j X_{\mathrm{k}}) / p(X_i X_j) \geq \gamma_{\min},$$
$$p(X_k / \neg(X_i\, X_j)) = p(\neg(X_i X_j) X_{\mathrm{k}}) / p(\neg(X_i\, X_j)) \geq \gamma_{\min},$$
$$p(\overline{X}_k / X_i\, X_j) = p(X_i X_j \overline{X}_{\mathrm{k}}) / p(X_i X_j) \geq \gamma_{\min},$$
$$p(\overline{X}_k / X_i\, X_j) = p(\neg(X_i X_j) \overline{X}_{\mathrm{k}}) / p(\neg(X_i\, X_j)) \geq \gamma_{\min}.$$

Если хотя бы одно из этих условий выполняется, то переход на п. 3.5. Иначе пара $< X_i\, X_j\, X_{\mathrm{k}}, p(X_i X_j X_{\mathrm{k}}) >$ из дальнейшего анализа исключается и выполняется переход на п. 3.1.

3.5. *Фильтрация 3*. Вычисляются значения коэффициентов регрессии:

$$R(X_k, X_i X_j) = [\, p(X_i X_j X_k) - p(X_i X_j)\; p(X_k)\,]/ \\ /\{\, p(X_i X_j)\,[1 - p(X_i X_j)\,]\}, \quad (23)$$

$$R(X_k, \overline{X}_i X_j) = [\, p(\overline{X}_i X_j X_k) - p(\overline{X}_i X_j)\; p(X_k)\,]/ \\ /\{\, p(\overline{X}_i X_j)[1 - p(\overline{X}_i X_j)]\}, \tag{24}$$

$$R(X_k, X_i \overline{X}_j) = [\, p(X_i \overline{X}_j X_k) - p(X_i \overline{X}_j)\; p(X_k)\,]/ \\ /\{\, p(X_i \overline{X}_j)[1 - p(X_i \overline{X}_j)]\}, \tag{25}$$

$$R(X_k, \overline{X}_i \overline{X}_j) = [\, p(\overline{X}_i \overline{X}_j X_k) - \\ p(\overline{X}_i \overline{X}_j)\; p(X_k)\,]/\{\, p(\overline{X}_i \overline{X}_j)[1 - p(\overline{X}_i \overline{X}_j)]\}, \tag{26}$$

$$R(\overline{X}_k, X_i X_j) = [\, p(X_i X_j \overline{X}_k) - p(X_i X_j)\; p(\overline{X}_k)\,]/ \\ /\{\, p(X_i X_j)[1 - p(X_i X_j)]\} = - R(X_k, X_i X_j) \tag{27}$$

$$R(\overline{X}_k, \overline{X}_i X_j) = [\, p(\overline{X}_i X_j \overline{X}_k) - p(\overline{X}_i X_j)\; p(\overline{X}_k)\,]/ \\ \{\, p(\overline{X}_i X_j)[1 - p(\overline{X}_i X_j)]\} = - R(X_k, \overline{X}_i X_j), \tag{28}$$

$$R(\overline{X}_k, X_i \overline{X}_j) = [\, p(X_i \overline{X}_j \overline{X}_k) - p(X_i \overline{X}_j)\; p(\overline{X}_k)\,]/ \\ /\{\, p(X_i \overline{X}_j)[1 - p(X_i \overline{X}_j)]\} = - R(X_k, X_i \overline{X}_j), \tag{29}$$

$$R(\overline{X}_k, \overline{X}_i \overline{X}_j) = [\, p(\overline{X}_i \overline{X}_j \overline{X}_k) - p(\overline{X}_i \overline{X}_j)\; p(\overline{X}_k)\,]/ \\ /\{\, p(\overline{X}_i \overline{X}_j)[1 - p(\overline{X}_i \overline{X}_j)]\} = - R(X_k, \overline{X}_i \overline{X}_j), \tag{30}$$

$$R(X_i X_j, X_k) = [\, p(X_i X_j X_k) - p(X_i X_j)\; p(X_k)\,]/ \\ /\{\, p(X_k)[1 - p(X_k)]\} = - R(\neg (X_i X_j), X_k), \tag{31}$$

$$R(\overline{X}_i X_j, X_k) = [\, p(\overline{X}_i X_j X_k) - p(\overline{X}_i X_j)\; p(X_k)\,]/ \\ /\{\, p(X_k)[1 - p(X_k)]\} = - R(\neg (\overline{X}_i X_j), X_k), \tag{32}$$

$$R(X_i \overline{X}_j, X_k) = [\, p(X_i \overline{X}_j X_k) - p(X_i \overline{X}_j)\; p(X_k)\,]/ \\ /\{\, p(X_k)[1 - p(X_k)]\} = - R(\neg (X_i \overline{X}_j), X_k), \tag{33}$$

$$R(\overline{X}_i \overline{X}_j, X_k) = [\, p(\overline{X}_i \overline{X}_j X_k) - p(\overline{X}_i \overline{X}_j)\; p(X_k)\,]/ \\ /\{\, p(X_k)[1 - p(X_k)]\} = - R(\neg (\overline{X}_i \overline{X}_j), X_k), \tag{34}$$

$$R(X_i X_j, \overline{X}_k) = [\, p(X_i X_j X_k) - p(X_i X_j)\; p(\overline{X}_k)\,]/ \tag{35}$$

$$/\{\, p(X_k)\,[1 - \, p(X_k)\,]\} = - \, R(\neg\,(X_i X_j), \overline{X}_k)\,,$$

$$R(\overline{X}_i X_j, \overline{X}_k) = [\, p(\overline{X}_i X_j X_k) \, - \, p(\overline{X}_i X_j)\; p(\overline{X}_k)\,]/$$

$$/\{\, p(X_k)\,[1 - \, p(X_k)\,]\} = - \, R(\neg(\overline{X}_i X_j), \overline{X}_k)\,, \qquad (36)$$

$$R(X_i \overline{X}_j, \overline{X}_k) = [\; p(X_i \overline{X}_j X_k) - \, p(X_i \overline{X}_j)\; p(\overline{X}_k)\,]/$$

$$/\{\, p(X_k)\,[1 - \, p(X_k)\,]\} = - \, R(\neg\,(X_i \overline{X}_j), \overline{X}_k)\,, \qquad (37)$$

$$R(\overline{X}_i \overline{X}_j, X_k) = [\, p(\overline{X}_i \overline{X}_j X_k) \, - \, p(\overline{X}_i \overline{X}_j)\; p(\overline{X}_k)\,]/$$

$$/\{\, p(X_k)\,[1 - \, p(X_k)\,]\} = - R(\neg\,(\overline{X}_i \overline{X}_j), X_k)\,. \qquad (38)$$

Если находится хотя бы один вариант коэффициента регрессии, в котором условие 3 постановки задачи выполнено, то трехлитерный узел $< X_i \; X_j \; X_{\mathrm{k}} >$ остается в качества кандидата на построение более длинных ассоциативных правил (либо за счет удлинения посылки, либо за счет удлинения следствия, либо за счет обеих компонент ассоциативного правила).

В этом случае из множества регрессий, вычисленных по формулам (23)–(38) выбираются те, которые удовлетворяют условию 3 в постановке задач. Они в дальнейшем могут использоваться для генерации соответствующих ассоциативных правил вида $X \to Y$, где $Y$ – первый аргумент коэффициента регрессии, а $X$ – ее второй аргумент.

Если не находится ни одного варианта регрессионного правила, удовлетворяющего указанному условию, то этот узел далее в АБС не рассматривается (не участвует в дальнейшем построении АБС).

Переход на п. 3.1.

4. *Алгоритм построения и фильтрации паттернов четвертого порядка*

Алгоритм построения и фильтрации узлов АБС четвертого порядка полностью аналогичен ранее рассмотренным алгоритмам для узлов второго и третьего порядков. Различие состоит только в используемых формулах, которых в случае узлов четвертого порядка больше, и они несколько сложнее. Описание данного алгоритма можно найти на веб-сайте по адресу *http://space.iias.spb.su/ai/abn/*.

На практике обычно можно ограничиться построением ассоциативных правил с двухлитерными посылками, поскольку известно, что связи, которые содержат более двух–трех переменных,

скорее всего, статистически недостоверны из-за ограниченности экспериментальных данных. Кроме того, хорошо известно, что человек в состоянии осмыслить и интерпретировать связи, которые содержат не более трех переменных. Поэтому в большинстве случаев можно ограничиться поиском ассоциативных правил, порождаемых узлами АБС не более чем третьего порядка.

**6. Предобработка данных, поиск правил в задаче обучения классификации и экспериментальные результаты.**

**6.1. Постановка задачи и предобработка данных для построения агрегатов.** Одним из важных приложений описанного алгоритма поиска ассоциаций является поиск правил в задаче *обучения* классификации. Для этого не требуется каких-либо модификаций этого алгоритма. В случае бинарной классификации достаточно только ввести дополнительный узел в АБС, который будет соответствовать метке класса. В случае, когда рассматривается задача классификации с большим числом классов, можно рекомендовать один из известных подходов редукции такой задачи к нескольким задачам классификации. Наиболее распространены, как известно, три подхода:

в одном из них строятся классификаторы, которые имеют целью попарное разделение классов с последующей обработкой полученных решений для принятия итогового решения;

в другом варианте для каждого класса строится бинарный классификатор, который принимает решение либо в пользу своего класса, либо против, т. е. классификатор "голосует" в пользу множества всех остальных классов, не указывая конкретно альтернативный класс;

в третьем случае рассматривается дерево классификации (чаще – бинарное), в котором листья помечены именем одного из классов, и общее число листьев равно числу классов.

В любом из этих подходов в режиме обучения может быть использован предложенный алгоритм поиска правил в виде "*Если…, то…*". В данном разделе будет рассмотрен пример поиска правил в задаче множественной классификации, когда используется второй из упомянутых подходов, однако базовые множества (унарные предикаты принадлежности) формируются из условия разделения пар классов. Далее эта особенность поясняется на примере.

Рассмотрим в качестве приложения задачу предсказания рейтинга фильма, который дается некоторым зрителем, когда в качестве исходной информации используется год выпуска фильма и его жанр. В качестве обучающих данных воспользуемся базой данных *MovieLense*

[26]. Этот пример достаточно прост, но в нем исходное пространство принятия решений включает в себя признаки числового и категориального типа, что характерно для большой части приложений в области классификации.

Хотя исходные данные для обучения представлены в реляционной базе данных несколькими таблицами, над которыми задана онтология, здесь не рассматривается вопрос о том, каким образом формируются запросы к базе данных, чтобы сформировать таблицу обучающих данных для отдельного зрителя и каждого из классов, поскольку этот аспект вне темы данной работы. Здесь предположим, что имеются плоские таблицы обучающих данных для каждого из классов решений конкретного зрителя, при этом полагаем, что зритель оценивает фильмы по 5-балльной шкале. Соответственно, каждая запись обучающих данных имеет метку из множества $\Omega=\{\omega_1, \omega_2, \ldots, \omega_5\}=\{1, 2, \ldots, 5\}$.

Общая идея предобработки данных, имеющей целью агрегирование отдельных значений признаков, в рассматриваемой задаче классификации[1] состоит в следующем. Пусть необходимо построить унарные предикаты принадлежности $B_i(\omega_k)$ такими, чтобы они, по возможности, могли играть роль первичных классификаторов. Более строго говоря, область истинности $X_i(\omega_k)$ унарного предиката $B_i(\omega_k)$ целесообразно строить таким образом, чтобы *для обучающего примера* $\boldsymbol{X}=\{x_1, \ldots, x_N, \omega_k\}$, если $x_r \in X_i$ (при этом унарный предикат $B_i$ истинен), то *для* $\forall\, \omega_l \in \Omega$ и $k \neq l$: $p(\omega_k / \boldsymbol{X}) > p(\omega_l / \boldsymbol{X})$. Содержательно в этом условии записано, что условная вероятность класса $\omega_k$ для примера $\boldsymbol{X}$ должна быть больше, чем условная вероятность любого другого класса. Если используется множественная классификация, в которой рассматривается подход с попарным разделением классов, тогда то же самое условие должно формулироваться для выбранной пары классов. Если не выполняется указанное неравенство, то и в этом случае соответствующий агрегат может оказаться полезным в правилах,

[1] Описываемый подход к построению агрегатов является общим при анализе ассоциаций, здесь он используется применительно к задаче классификации.

включающих его в комбинации с другими агрегированными признаками. В любом случае нужно стремиться к максимизации условной вероятности $p(\omega_k / \boldsymbol{X})$.

Описанная идея может быть достаточно просто реализована алгоритмически. На основании обучающей выборки класса $\omega_k$ для каждого значения (первичного) признака $x_r$ можно оценить вероятность $p(x_i / \omega_k)$. Тогда алгоритм формирования множества $X_i(\omega_k)$ (например, для задачи множественной классификации с попарным разделением классов, что используется далее в примере) сводится к простому правилу:

*Правило 1.* Если $p(x_i/\omega_k) \geq p(x_i/\omega_l) - \Delta$, то $x_i \in X_i(\omega_k)$. В противном случае значение $x_i \notin X_i(\omega_k)$. Здесь $\Delta$ – положительное число, играющее роль порога, выбираемого в конкретной задаче индивидуально.

Это правило целесообразно использовать в случае, когда выборка обучающих данных либо мала, либо не вполне представительна. Это имеет место, например, когда число обучающих примеров в разных классах не соответствует априорным вероятностям классов [7]. Если же выборка достаточно велика и представительна, и по ней можно достаточно достоверно оценить априорные вероятности классов, то в этом случае целесообразно использовать байесовские оценки апостериорных вероятностей классов. Сформулируем это в виде правила 2.

*Правило 2.* Если $p(\omega_k / x_i) > p(\omega_l / x_i) + \Delta$ ($\Delta$–положительное число), то значение $x_i$ признака $X$ включается в множество $X_i(\omega_k)$.

Описанный выше подход ориентирован на построение агрегатов на категориальных признаках. Аналогичный подход может использоваться и для числовых признаков, если непрерывные множества значений таких признаков заменить дискретным множеством представителей интервалов.

Другой подход, который использовался в описанном далее примере для числовых признаков, базируется на более традиционном методе разделения непрерывного интервала значений признака на два интервала путем оптимизации некоторой меры информативности. В

описанном далее примере для этих целей использовался критерий информационного выигрыша, предложенный Дж.Р.Квинланом в его методе обучения, известном под названием C4.5 [24].

**6.2. Экспериментальные результаты.** Рассматривается пример с двумя *базовыми* признаками:

первый из них – это *жанр* фильма, *Genre,* который принимает значения из множества $\boldsymbol{G}$ ={*Action, Adult, Adventure, Animation, Biography, Crime, Comedy, Documentary, Family, Fantasy, Horror, Music, Musical, Mystery, Sci-Fi, Short, Sport, Thriller, War*};

второй – это год выпуска фильма в США, *ReleaseYear*, $\boldsymbol{J}$ =[*1976;2005*].

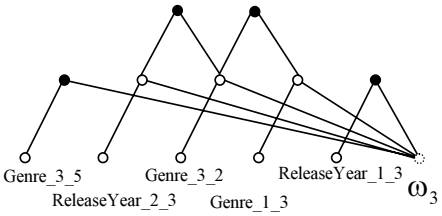


Рис. 5. АБС для класса $\omega_3$.

Имеется всего пять вариантов оценки фильма $\Omega$ = {1,2, … ,5}, причем рейтинг "1" соответствует самой низкой оценке. Для выбранного зрителя объемы обучающих выборок в базе данных *MovieLens* [26] для каждого из классов таковы: $\boldsymbol{N_1}$ = *2,* $\boldsymbol{N_2}$ = *6,* $\boldsymbol{N_3}$ = *50,* $\boldsymbol{N_4}$ = *86,* $\boldsymbol{N_5}$ = *25*. Всего, таким образом, обучающая выборка содержит 169 примеров.

По правилу 1 (см раздел 5.3.1) построены агрегаты для всех классов для признака *жанр*, при этом они строились для всех пар классов $X_l(\omega_k, \omega_l)$, где первый аргумент указывает целевой класс, а второй – альтернативный класс. Для первичного признака *жанр* для каждого класса таким образом было построено по четыре агрегата. Например, для класса $\omega_3$ эти агрегаты получились такими:

$X_1(\omega_3, \omega_1)$ = {*Action, Adult, Adventure, Animation, Biography, Crime, Documentary, Family, Fantasy, Music, Musical, Mystery, Sci-Fi, Short, Sport, Thriller*},

$X_1(\omega_3, \omega_2)$ = {*Adult, Adventure, Animation, Biography, Documentary, Fantasy, Music, Musical, Mystery, Sci-Fi, Short, Sport*},

$X_1(\omega_3, \omega_4)$ = {*Adult, Music*},

$X_1(\omega_3, \omega_5) = \{Adult,\ Musical,\ Sci\text{-}Fi\}$,

$X_1(\omega_1, \omega_3) = \{Comedy\}$,

$X_1(\omega_2, \omega_3) = \{Horror\}$,

$X_1(\omega_4, \omega_3) = \{History,\ Horror,\ War\}$,

$X_1(\omega_5, \omega_3) = \{History,\ War\}$.

То же самое было проделано для признака *ReleaseYear*. Для класса $\omega_3$, например, построены следующие агрегаты:

$X_1(\omega_3,\ \omega_1) = \{J \geq 1992{,}5\},\ X_1(\omega_3,\ \omega_2) = \{J \geq 1996{,}5\}$,

$X_1(\omega_3, \omega_4) = \{J \geq 1983{,}5\},\ X_1(\omega_3,\ \omega_5) = \{J \geq 1980\}$.

Для поиска правил типа "*Если..., то…*" в каждом классе строилось по две АБС, одна из которых была предназначена для поиска правил "в пользу целевого класса", а другая – для поиска правил в пользу альтернативного множества классов. Такие правила принято называть *запретами*. Каждая из этих АБС в начале построения правил включала по 9 базовых узлов, одному из которых ставилась в соответствие метка целевого класса.

Для построения АБС, в которой *минимальные узлы* соответствуют посылкам правил в пользу целевого класса, заключению соответствовала метка класса. В правилах в пользу альтернативного класса в качестве заключения использовалось отрицание истинности переменной, которая ставилась в соответствие метке класса. Заметим, что этот случай соответствует поиску негативных ассоциативных правил. Их поиск выполняется здесь в рамках той же самой модели, что и поиск положительных ассоциаций.

Еще раз отметим, что *минимальные узлы* каждой из АБС

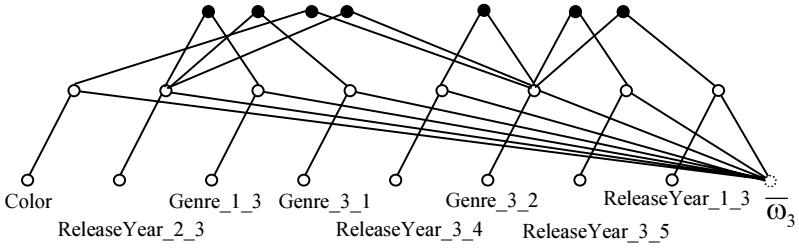


Рис. 6. АБС для класса $\overline{\omega}_3$.

соответствуют посылкам правил в пользу целевого класса и в пользу альтернативного класса соответственно. Всего на основе первой АБС было отобрано 4 правила. Соответствующее число для АБС в пользу альтернативного класса равно 14. Примеры АБС, построенных таким образом для класса $\omega_4$, представлены на рис.5 и 6. На рис.6 представлена АБС, в которую включено только несколько наиболее сильных правил.

Найденные множества правил для каждого из классов могут быть далее использованы уже для построения механизмов классификации, с помощью того или иного метода объединения решений [7]. Однако эта задача уже выпадает из основного содержания данной работы.

**7. Заключение.** Предложенный в работе алгоритм базируется на чисто вероятностной трактовке задачи поиска ассоциативных правил и причинных закономерностей. В работе описана новая модель вероятностного пространства, которая рассчитана на использование неполного описания вероятностной модели предметной области. Для этих целей предложена модель недоопределенного вероятностного пространства, задающего некоторый класс распределений, которому принадлежит вероятностная модель рассматриваемого приложения. Для предложенной модели вероятностного пространства легко строится изоморфная ему вероятностная логика, которая позволяет, что очень важно, вычислять вероятностную меру любой логической формулы, представляющей некоторые закономерности предметной области путем вычисления вероятности случайного события, которое соответствует этой формуле по введенному изоморфизму в вероятностном пространстве. То же самое касается и логических формул, которые имеют свой изоморфный образ в недоопределенном вероятностном пространстве.

Структура для представления знаний в вероятностном и в недоопределенном вероятностном пространствах, которая в работе называется *ассоциативной байесовской сетью*, является удобной формой представления знаний с неопределенностью. Данная структура представляет минимальный объем информации, с помощью которой можно вычислить вероятности любых других событий (и логических формул), которые выражаются через заданные события (формулы) с помощью операций булевой алгебры событий (булевой алгебры формул – в логической форме АБС). Описанная логико – вероятностная модель в данной работе используется для решения одной из важных задач современной проблемы интеллектуального

анализа данных – поиска ассоциативных правил общего вида и причинных закономерностей.

С точки зрения сложности поиска ассоциативных правил, предложенный в работе алгоритм внешне выглядит не более эффективным, чем большинство традиционных алгоритмов их поиска, но модель АБС обладает определнными ресурсами для снижения объема вычислений, которые в данной работе не анализируются. Однако следует обратить внимание на то, что даже при объеме вычислений, сравнимом с тем, что требуют традиционные алгоритмы, например, алгоритмы типа Apriori, в предложенном алгоритме решается задача гораздо большего объема. Действительно, алгоритм, использующий АБС, рассчитан на поиск правил как с большим значением поддержки, так и правил, которые отвечают редким, но сильным закономерностям. А такая задача сама по себе намного сложнее, чем традиционная. Кроме того, при сравнимом объеме вычислений предложенный алгоритм позволяет находить негативные правила для любого варианта размещения отрицания на литерах, формирующих ассоциативное правило. А эта задача рассматривается как одна из весьма проблемных, но очень важных для практики. Наконец, очень важное свойство предложенного подхода и реализующего его алгоритма состоит в том, что он позволяет находить причинные связи, а не просто ассоциативные.

Описанный в работе алгоритм реализован программно практически в полном объеме. Он применялся для решения нескольких задач различной размерности. В частности, одно из приложений включало в себя тысячи признаков, причем и числовых, и категориальных. Благодаря использованию алгоритма агрегирования отдельных признаков (в результате построены признаки – агрегаты типа тех, что в работе обозначались символами $X_i$, $i = 1, \ldots, n$, которым соответствовали также унарные предикаты принадлежности; последние далее использовались для логического представления правил в причинно–следственной форме) общее число агрегатов было сокращено до сотен. Таким образом, АБС, построенная в этом приложении, содержала сотни базовых узлов. С ее помощью извлечено несколько сотен правил, которые далее (после редукции их множества по косвенно оцениваемым мерам информативности) весьма успешно использовались для решения задачи классификации данных на несколько классов. Объем тренировочных и тестовых данных составил (в разных вариантах) от десятков до нескольких сотен примеров. Отметим, что определенное время, причем немалое, использовалось

для извлечения примеров из реляционной базы данных, содержащей порядка 20 плоских таблиц и миллионы записей, в режиме *on–line* в контексте онтологии предметной области, содержащей десятки понятий. Тем не менее, даже на не самом мощном персональном компьютере время решения задачи оказалось в разумных пределах. Это позволяет оптимистически смотреть на прикладные возможности рассмотренной модели недоопределенного вероятностного пространств, его представления в терминах структуры АБС, а также на перспективы ее использования в области интеллектуальной обработки данных не только в приведенном классе задач, но и в ряде других.

**Благодарности.** Данная работа выполнялась при частичной поддержке гранта программы Отделения нанотехнологий и информационных технологий РАН (Программа "Информационные технологии и методы анализа сложных систем", направление № 1 "Системы автоматизации, обработки информации и поддержки принятия решений", проект № 1.12.)

## Литература

1. *Биркгоф Г., Барти Т.* Современная прикладная алгебра., М.: Мир, 1976. 370 с.
2. *Городецкий В. И.* Адаптация в экспертных системах// Изв. РАН Техническая кибернетика, 1993. № 5. С. 101-110.
3. *Городецкий В. И.* Алгебраические байесовские сети – новая парадигма экспертных систем// Юбилейный сб. тр. институтов Отделения информатики, вычислительной техники и автоматизации РАН. т.2. 1993.С. 120-141.
4. *Городецкий В. И. Тулупьев А.Л.* Формирование непротиворечивых баз знаний с неопределенностью//Изв. РАН, Теория и системы управления. 1997. № 5.
5. *Городецкий В. И.* Интервальные вероятностные меры неопределенности в инженерии знаний// Юбилейный сб. тр. СПИИРАН. СПб 1998.
6. *Городецкий В. И.* Моделирование недоопределенных знаний// Сб. докл. Междунар. конф. по мягким вычислениям (SCM'98). т.1. СПб. 1998. С. 98 – 102.
7. *Городецкий В. И., Серебряков С.В.* Методы и алгоритмы коллективного распознавания //Автоматика и Телемеханика. 2008. № 11. С. 3–40.
8. *Яглом И.М.* Булева структура и ее модели. М.: СовРадио. 1980. 192 С.
9. *Adamo J.-M.* Data Mining for Association Rules and Sequential Patterns. Springer, 2000.
10. *Agrawal R., Sricant R.* Fast Algorithm for Mining Association rules// Proc. of the 20th Intern. Conference on Very Large Databases. Santiago, Chile, 1994.
11. *Agrawal R, Imielinski T., Swami* A. Mining association rules between sets of items in large databases// Proc. of the ACM SIGMOD Conf. on Management of Data. Washington, D.C. 1993.
12. *Brin S., Motwani R., Silverstein C.* Beyond market baskets: generalizing association rules to correlations// Proc. of the ACM SIGMOD Intern. Conf. on Management of Data. 1997. P. 255–264.
13. *Dzeroski S.* Multi-relational data mining: An introduction// ACM SIGKDD Explorations Newsletter. Vol. 5, Issue 1. 2003. P. 1–16.
14. *S.Dzeroski , N.Lavrac* . Relational Data Mining. Springer, 2001.

15. *Fagin R., Halpern J. Y., Megiddo N. A.* Logic for Reasoning about Probabilities// Proc. of 3th IEEE Symposium on Logic and Computer Science. 1988.
16. *Fagin R., Halpern J. Y.* Uncertainty, Belief, and Probability// Proc. of 11th Intern. Joint Conf. on Artificial Intelligence. Detroit, Michigan, USA, 1989. P. 1161–1167.
17. *Gorodetski V. I.* Adaptation Problems in Expert Systems// Intern. J. of Adaptive Control and Signal Processing. 1992. Vol.6. P. 201-209.
18. *Halpern J. Y.* Reasoning about uncertainty. MIT Press: Cambridge. 2003.
19. *Han J., Fu Y., Wang W., Koperski K., Zaiane O*. DMQL: A Data Mining Query Language for Relational Databases// SIGMOD DMKD Workshop, 1996.
20. *Han J., Kamber M.* Data Mining: Concept and Techniques. Morgan Kaufman, 2000.
21. *Han J., Pei J., Yin Y.* Mining frequent patterns without candidate generation// Proc. of the ACM SIGMOD Intern. Conf. on Management of Data. 2000. P. 1–12.
22. *Liu H., Lu H, Feng L., Hussain F.* Efficient search of reliable exceptions// Proc. of the 3d Pacific Asia Conference on Knowledge Discovery and Data Mining, 1999. P. 194–204.
23. *Morzy M.* Efficient Mining of Dissociation Rules// Lecture Notes in Computer Sci 2006. Vol. 4081, 2006, Springer, P. 228–237.
24. Quinlan *J.* C4.5: Programs for Machine Learning// Elsevier Science & Technology Books, 1992. 316 P.
25. *Piatetsky–Shapiro G.* Discovery, analysis and presentation of strong rules// Knowledge Discovery in Databases. AAAI Press/MIT Press. 1991. P. 229–248.
26. *Movylence*. http://www.grouplens.org/node/73.
27. *Szathmary L., Napoli A., Valtchev P*. Towards Rare Itemset Mining// Proc. of the 19th IEEE Intern. Conf. on Tools with Artificial Intelligence (ICTAI '07). Patras, Greece, Oct 2007. P. 305-312.
28. *Shortliffe E.* Computer–based medical consultations: MYSYN. N. Y.: Elsevier. 1976.
29. *Tze A., Sim H., Zutshi S., Indrawan M., et al.* The Discovery of Coherent Rules. IAENG Intern. J. of Computer Sci. 2008. Vol. 35, № 3.
30. *Wu X., Zhang C., Zhang S.* Efficient Mining of Both Positive and Negative Association Rules// ACM Trans. on Information Systems. 2004. Vol. 22, № 3.P. 381–405.
31. *Zhang C., Zhang S.* Association rule mining. Springer, 2002. 240 P.
32. *Web site of LIS*. http://spaec.iias.spb.su/ai/abn/

**Городецкий Владимир Иванович**, Главный научный сотрудник лаборатории интеллектуальных систем Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Окончил Ленинградскую военно-воздушную инженерную академию им. А.Ф.Можайского в 1960 г. и математико-механический факультет Ленинградского государственного университета в 1970 г. Д. техн. наук, проф.; Заслуженный деятель науки Российской Федерации. Опубликовал более 350 работ, 9 монографий и учебных пособий. Область научных интересов: теория и технология многоагентных систем, инструментальные средства многоагентных систем, многоагентные приложения в области защиты компьютерных сетей, объединения данных из гетерогенных источников, оценки ситуаций, управления воздушным движением, методы и средства интеллектуальной обработки данных и извлечения знаний, методы распределенного обнаружения знаний в данных, peer–to–peer системы, распределенное принятие решений. Адрес: gor@iias.spb.su, СПИИРАН, 14-я линия В. О., д. 39, Санкт-Петербург, 199178, РФ; раб.тел. +7(812)323-3570, факс +7(812)328-4450, http://space.iias.spb.su/webarchive/ai/gorodetsky/index.jsp.htm.

**Vladimir Gorodetsky**, Prof. of Computer Science, Chief Scientist of The Intelligent Systems Laboratory of The St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science. Graduated from the Military Air Force Engineer Academy in

St. Petersburg (1960) and Mathematical and Mechanical Department of the St. Petersburg State University (1970), Ph.D. (1967) and Doctor of Technical Sciences (1973). Main publications (more 350) are related to the areas Artificial intelligence, in particular, Distributed intelligence, Knowledge-based planning and scheduling, Pattern and knowledge discovery in databases, P2P data mining and knowledge discovery, Data and information fusion, Decision combining, Distributed classification, Multi-agent system technology and software tools, P2P agent systems and P2P Agent platform, Multi-agent applications, in particular, Computer network security, Air traffic control, Project management.
Address gor@iias.spb.su, SPIIRAS, 39, 14 -th Line V.O., St. Petersburg, 199178, Russia;
office phone +7(812)323-3570, fax +7(812)328-4450,
http://space.iias.spb.su/webarchive/ai/gorodetsky/index.jsp.htm

**Самойлов Владимир Владимирович,** научный сотрудник Санкт-Петербургского институт информатики и автоматизации РАН. Окончил Высшее военно-морское инженерное училище им. Ф.Э. Дзержинского в 1993 г. Опубликовал более 30 работ. Область научных интересов: многоагентные системы, байесовские сети, системы объединения данных из разных источников, методы распределенного обучения, базы данных, программирование. Адрес: samovl@iias.spb.su, СПИИРАН, 14-я линия В. О., д. 39, Санкт-Петербург, 199178, РФ; раб.тел. +7(812)323-3570, факс +7(812)328-4450.

**Vladimir Samoylov,** research fellow of The Intelligent Systems Laboratory of The St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science. Graduated from the Naval Engineering College after F. Dzerzhinsky (1993). Main publications (more 30) are related to the areas Artificial intelligence, in particular, Distributed Intelligence, Data mining and knowledge discovery, Bayesian networks, Multi-agent system technology and software tools, P2P agent platform, Service-oriented multi-agent systems, Data bases, Software engineering, Multi-agent applications, in particular, Computer network security, Air traffic control, and Project management.
Address: samovl@iias.spb.su, SPIIRAS, 39, 14 -th Line V.O., St. Petersburg, 199178, Russia;
office phone +7(812)323-3570, fax +7(812)328-4450.