Д.В. Виноградов

ВКФ-метод[1] интеллектуального анализа данных: обзор результатов и открытых проблем

Аннотация. Статья содержит описание текущего состояния дел с исследованием ВКФ-метода интеллектуального анализа данных. Этот метод соединяет в себе три когнитивных процедуры (индукции, абдукции и аналогии), основываясь на вероятностном алгоритме поиска сходств. Сформулированы основные известные результаты и открытые проблемы.

Ключевые слова: сходство, цепь Маркова, ВКФ-кандидат, контр-пример, предсказание по аналогии.

Введение

Первым человеком, который ввел и начал систематически изучать когнитивные процедуры, основанные на операции сходства, был проф. Виктор Константинович Финн [15]. Начиная с 1981 г., он и его ученики развивают так называемый ДСМ-метод автоматического порождения гипотез [8]. ДСМ-метод назван в честь известного английского философа, экономиста и логика Джона Стьюарта Милля. Используя технику многозначных логик, В.К. Финну с коллегами [1, 2] удалось поставить систему индуктивной логики Милля [11] на четкие логические основания. Ключевой компонентой этого подхода является бинарная операция сходства [7].

Примерно в это же самое время группа зарубежных исследователей под руководством проф. Рудольфа Вилле, основываясь на теории решеток, разработала теорию формальных понятий (ТФП) [20]. Эта теория оказалась полезной ДСМ-методу тем, что многие алгоритмические конструкции могут быть заимствованы из ТФП для построения решетки всех сходств. Правда, один из самых эффективных алгоритмов вычисления всех сходств – алгоритм «Замыкай-по-одному» [10] – был первоначально создан внутри ДСМ-сообщества и лишь затем переведен на язык ТФП.

Второй когнитивной процедурой стало предсказание по аналогии, что превратило ДСМ-метод в средство интеллектуального анализа данных [17].

Третья когнитивная процедура – абдуктивное принятие гипотез – возникло в трудах В.К. Финна в результате осмысления наследия известного американского математика и логика Чарльза Сэндерса Пирса [12].

После выяснения сути указанных когнитивных процедур отечественные ученые под руководством В.К. Финна смогли создать единую систему, объединяющую все эти процедуры в одно целое. Эта система и получила название ДСМ-метод [16]. В последнее время В.К. Финн и М.А. Михеенкова [19] создали еще один синтез индукции, абдукции и аналогии на основании булевско-алгебраического понятия сходства.

Следует признать, что имеются некоторые сложности в применении ДСМ-метода к анализу данных. Во-первых, множество порождаемых ДСМ-гипотез может оказаться экспоненциально велико по сравнению с размером

[1]Назван так в честь проф. Виктора Константиновича Финна, идеи которого он содержит.

обучающей выборки. В реальных экспериментах это проявляется как значительное время вычислений, в результате которых порождается огромное число гипотез. Правда, похожие гипотезы действуют одинаково (когда одновременно применимы). Во-вторых, С.О. Кузнецовым [9], М.И. Забежайло и др. были доказаны пессимистические оценки сложности для многих ДСМ-процедур (так называемые **NP**-полнота и **#P**-полнота). В-третьих, попытки присоединить дедуктивный вывод столкнулись с проблемами. Так, Д.П. Скворцовым [13] установлена неарифметичность стандартного подхода через кванторы по конечным множествам, а автором этой статьи [4] – невыразимость этой теории средствами логики предикатов первого порядка.

Наконец, автор [6] сумел обнаружить еще одну трудность: существование так называемых случайных ДСМ-гипотез, которые возникают тогда, когда вычисляется сходство двух (или более) обучающих примеров, каждый из них имеет свой механизм порождения целевого свойства. Это сходство оказывается фрагментом (набором общих признаков), случайно имеющемся в каждом из этих объектов. Возникновение таких сходств следует рассматривать как аналог «переобучения» для многих методов машинного обучения, когда максимальный учет информации из обучающей выборки приводит к модели, демонстрирующей плохую предсказательную способность. Естественным способом борьбы с этим феноменом может быть правильный выбор признаков, описывающих объекты. Однако этот вопрос нуждается в дополнительных исследованиях.

Для преодоления всех указанных трудностей автором был предложен вероятностно-комбинаторный подход. Так как некоторые ингредиенты заимствованы автором из теории формальных понятий, назовем его вероятностно-комбинаторным формальным методом, сокращенно ВКФ-метод.

1. Базовые понятия

Сходство – это бинарная операция, задающая структуру нижней полурешетки с наименьшим элементом.

Формальный контекст можно описать как бинарное отношение между элементами множества *O*, которые мы называем *именами объектов*, и элементами множества *F*, которые мы называем *признаками*. Если в строке, соответствующей объекту $o \in O$, и столбце, соответствующим фрагменту $f \in F$, стоит единица, то мы говорим, что *объект o обладает признаком f*, и обозначаем это через *oIf*. В противном случае, говорим, что *объект o не имеет признака f*.

Для подмножества $A \subseteq O$ объектов его *сходством* называется подмножество $A' = \{f \in F: \forall o \in A[oIf]\} \subseteq F$. Полагаем $\emptyset' = F$. На самом деле, это определение совпадает с последовательным вычислением побитового умножения строк, соответствующих отобранным во множество *A* объектов.

Для подмножества $B \subseteq F$ признаков его *сходством* называется подмножество $B' = \{o \in O: \forall f \in B[oIf]\} \subseteq O$. Полагаем $\emptyset' = O$.

Определение 1. Пару $\langle A, B\rangle$ назовем *ВКФ-кандидатом*, если $A = B' \subseteq O$ и $B = A' \subseteq F$.

Определение 2. Операция *замыкай-по-одному-вниз* на ВКФ-кандидате $\langle A, B\rangle$ и объекте $o \in O$ порождает пару $CbODown(\langle A, B\rangle, o) = \langle (A \cup \{o\})'', (A \cup \{o\})'\rangle$. Операция *замыкай-по-одному-вверх* на ВКФ-кандидате $\langle A, B\rangle$ и признаке $f \in F$ порождает пару $CbOUp(\langle A, B\rangle, f) = \langle (B \cup \{f\})', (B \cup \{f\})''\rangle$.

2. Основные алгоритмы

Если некоторые сходства заведомо плохи (случайные), а огромное большинство сходств предсказывает по аналогии примеры одинаковым образом, то нет никакой необходимости вычислять их все, достаточно найти случайное подмножество сходств. Автор развил вероятностно-комбинаторный формальный метод (ВКФ-метод) для реализации этой идеи.

ВКФ-метод использует синтез познавательных процедур из ДСМ-метода, модифицируя их:

- индуктивное обобщение данных;
- абдуктивное уточнение и принятие гипотез (порождая дополнительные гипотезы для объяснения обучающих примеров);
- предсказание целевого свойства по аналогии с обучающими примерами.

Здесь мы дополнили абдукцию - условие принятия порожденных гипотез - процедурой абдуктивного уточнения множества гипотез. Абдуктивное уточнение заключается в применении операции $CbODown(\langle A, B\rangle, o)$ к каждому исходному обучающему примеру *o* и каждой порожденной на шаге индукции ВКФ-

гипотезе $\langle A, B\rangle$. Это было необходимо для увеличения шансов найти ВКФ-гипотезу, которая правильно объясняет исходный обучающий пример. Из-за вероятностного характера порождения гипотез все ВКФ-гипотезы, включающиеся в выбранный пример, могли быть пропущены.

Так как абдуктивное уточнение увеличивает множество гипотез, мы перенесли ее на второй шаг, чтобы предсказание по аналогии работало с большими шансами на успех. Прежде всего, нам необходимо представить вероятностные алгоритмы для нахождения сходств:

Data: множество обучающих (+)-примеров; внешние функции $CbOUp(\ ,\ )$ и $CbODown(\ ,\ )$ операций «замыкай-по-одному»
Result: ВКФ-кандидат $\langle A, B\rangle$
O := (+)-примеры, F := признаки; $I \subseteq O \times F$ – формальный контекст для (+)-примеров;
A :=O; B = O'; i := 0;
while (i < T) **do**
$R := (O \backslash A) \cup (F \backslash B)$;
Выбираем случайный элемент $r \in R$;
if $(r \in O \backslash A)$ **then**
$\langle A, B\rangle := CbODown(\langle A, B\rangle, r)$;
end
else
$\langle A, B\rangle := CbOUp(\langle A, B\rangle, r)$;
end
end

Алгоритм 1: Немонотонная цепь Маркова

Теорема 1. *Алгоритм 1 соответствует цепи Маркова.*

Data: множество обучающих (+)-примеров; внешние функции $CbOUp(\ ,\ )$ и $CbODown(\ ,\ )$ операций «замыкай-по-одному»
Result: ВКФ-кандидат $\langle A, B\rangle$
O := (+)-примеры, F := признаки; $I \subseteq O \times F$ – формальный контекст для (+)-примеров;
A :=O; B = O'; $R := O \cup F$; i := 0;
while (i < T) **do**
Выбираем случайный элемент $r \in R$;
if $(r \in O)$ **then**
$\langle A, B\rangle := CbODown(\langle A, B\rangle, r)$;
end
else
$\langle A, B\rangle := CbOUp(\langle A, B\rangle, r)$;
end
end

Алгоритм 2: Монотонная цепь Маркова

Теорема 2. *Алгоритм 2 соответствует цепи Маркова.*

Data: множество обучающих (+)-примеров; внешние функции $CbOUp(\ ,\ )$ и $CbODown(\ ,\ )$ операций «замыкай-по-одному»
Result: ВКФ-кандидат $\langle A, B\rangle$
O := (+)-примеры, F := признаки; $I \subseteq O \times F$ – формальный контекст для (+)-примеров;
$R := O \cup F$; $Min := \langle O, O'\rangle$; $Max := \langle F', F\rangle$;
while $(Min \neq Max)$ **do**
Выбираем случайный элемент $r \in R$;
if $(r \in O)$ **then**

$Min := CbODown(Min, r)$; $Max := CbODown(Max, r)$;
end
else
$Min := CbOUp(Min, r)$; $Max := CbOUp(Max, r)$;
end
end

Алгоритм 3: Спаривающая цепь Маркова

Заметим, что состоянием изменяемых переменных в цикле (= состоянием спаривающей цепи Маркова) является упорядоченная пара ВКФ-объектов $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$. Первоначально меньший ВКФ-кандидат совпадает с наименьшим ВКФ-кандидатом $Min := \langle O, O'\rangle$, а больший - с наибольшим $Max := \langle F', F\rangle$. В цикле к обоим ВКФ-кандидатам применяется одна и та же операция *CbODown* с выбранным объектом или *CbOUp* с выбранным признаком.

Процесс останавливается, когда меньший ВКФ-кандидат совпадет в большим. Тогда этот общий ВКФ-кандидат и выдается Алгоритмом 3.

Теорема 3. *Алгоритм 3 соответствует цепи Маркова.*

Определение 3. Состояние вида $\langle A, B\rangle = \langle A, B\rangle$ спаривающей цепи Маркова для совпадающей пары ВКФ-кандидатов называется *эргодическим*. Состояние вида $\langle A_1, B_1\rangle < \langle A_2, B_2\rangle$ – *невозвратным*.

Теорема 4. *Вероятность того, что состояние* $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$ *спаривающей цепи Маркова окажется невозвратным, стремится к нулю, когда* $t \to \infty$.

Так как спаривающая цепь Маркова может иметь траектории существенно разной длины, то возможно применение следующей техники

остановки длинной траектории и запуска цепи заново.

Определение 4. Если $T_1, \dots, T_r$ – независимые целочисленные случайные величины, имеющие распределение времени T склеивания Алгоритма 3, то *верхняя граница склеивания* по r предварительным прогонам определяется как $\hat{T} = T_1 + \cdots + T_r$.

На практике предлагается сделать r прогонов спаривающей цепи Маркова с соответствующими временами склеивания $t_1, \dots, t_r$ и взять оценку $t_1 + \cdots + t_r$ верхней границы склеивания. Оценим, как соотносятся вероятности $\mu(R)$ и $\mu_{\hat{T}}(R)$ попадания в множество R эргодических состояний для исходной и остановленной по верхней границе $\hat{T}$ склеивания спаривающих цепей Маркова.

Теорема 5. *Для любого* R $c\ \mu(R) = \rho$ *и* $r > \log_2\left(1 - \frac{1}{\rho}\right)$ *имеем* $\mu_{\hat{T}}(R) \geq \rho - \frac{1}{2^r - 1}$.

Индуктивное обобщение обучающих примеров осуществляется следующей процедурой:
Data: множество обучающих (+)- и (-)-примеров; число N порождаемых ВКФ-гипотез
Result: выборка S ВКФ-гипотез объема N
O := (+)-примеры, F := признаки; $I \subseteq O \times F$ – формальный контекст для (+)-примеров;
C := (-)-примеры, $S := \emptyset$; i := 0;
while (i < N) **do**
    Породить ВКФ-кандидата $\langle A, B\rangle$ с помощью цепи Маркова;
    hasObstacle := **false**;
    for ($c \in C$) **do**
        if ($B \subseteq c'$) **then**
        hasObstacle := **true**;
        end
    end
    if (*hasObstacle* = **false**) **then**
        $S := S \cup \{\langle A, B\rangle\}$; i := i+1;
    end
end

Алгоритм 4: Процедура индуктивного обобщения

Проверка условия $B \subseteq c'$ в Алгоритме 4 означает, что фрагмент B ВКФ-кандидата $\langle A, B\rangle$ вкладывается во фрагмент (множество признаков) контр-примера c. Любое такое вложение означает, что ВКФ-кандидат нарушает условие «запрета на контр-пример». Если ВКФ-кандидат преодолевает все такие проверки, то он становится ВКФ-гипотезой (о причине наличия целевого свойства).

Вторая когнитивная процедура ВКФ-метода – алгоритм абдуктивного уточнения множества гипотез.
Data: выборка S ВКФ-гипотез, внешняя функция *CbODown*(,) операции «замыкай-по-одному-вниз»
Result: расширенная выборка S^+ ВКФ-гипотез
$S^+ := \emptyset$;
O := (+)-примеры, C := (-)-примеры;
for ($o \in O$ и $\langle A, B\rangle \in S$) **do**
    Вычислить ВКФ-кандидата $\langle X, Y\rangle := CbODown(\langle A, B\rangle, o)$;
    Explained(o) := **false**; *hasObstacle* := **false**;
    for ($c \in C$) **do**
        if ($Y \subseteq c'$) **then**
        hasObstacle := **true**;
        end
    end
    if (*hasObstacle* = **false**) **then**
        $S^+ := S^+ \cup \{\langle Y, Y\rangle\}$; *Explained(o)* := **true**;
    end
end

Алгоритм 5: Процедура абдуктивного уточнения

Как видно из Алгоритма 5 абдуктивное уточнение заключается в применении оператора *CbODown* к каждому исходному обучающему примеру и каждой порожденной на шаге индукции ВКФ-гипотезе.

Проверка условия $Y \subseteq c'$ в Алгоритме 5 означает, что фрагмент Y ВКФ-кандидата $\langle X, Y\rangle$ вкладывается во фрагмент (множество признаков) контр-примера c. Так проверяется условие «запрета на контр-пример».

Наконец, процедура предсказания целевого свойства по аналогии с обучающими примерами:
Data: расширенная выборка S^+ ВКФ-гипотез, файл (τ)-примеров
Result: предсказанные свойства (τ)-примеров
X := (τ)-примеры;
for ($o \in X$) **do**
    PredictPositively(o) := false;
    for ($\langle A, B\rangle \in S^+$) **do**
        if ($B \subseteq o'$) **then**
        PredictPositively(o) := **true**;
        end
    end
end

Алгоритм 6: Процедура предсказания по аналогии

Процедура предсказания по аналогии (Алгоритм 6) пытается найти вложение хотя бы одного фрагмента $B \subseteq o'$, соответствующего хотя бы одной из порожденных (индукцией и абдукцией) ВКФ-гипотез $\langle A, B\rangle$, в каждый (τ)-пример o. Если такое вложение случается, то для этого (τ)-примера предсказывается наличие целевого свойства по аналогии с родителями $A \subseteq O$ ВКФ-гипотезы $\langle A, B\rangle$, чей фрагмент B вложился. Иначе предсказывается отсутствие целевого свойства у этого (τ)-примера o.

Для выбора числа N запусков спаривающей цепи Маркова (Алгоритма 3) полезно применение следующей теоремы (мы используем объекты, представленные для предсказания).

Определение 5. *Нижнее полупространство* $H_\kappa^{\downarrow}(o)$, определяемое объектом o с фрагментом $o' \subseteq F$, задается линейным неравенством $x_{j_1} + \cdots + x_{j_k} < \kappa$, где $F \backslash o' = \{f_{j_1}, \ldots, f_{j_k}\}$ и $0 < \kappa < 1$.

Лемма 1. *Пример o предсказывается положительным тогда и только тогда, когда в любом его нижнем полупространстве содержится хотя одна ВКФ-гипотеза.*

Теперь мы будем проводить рассуждения, аналогичные рассуждениям В.Н. Вапника и А.Я. Червоненкиса [3], хотя нас будет интересовать только вероятность ошибки «первого рода» (отказ от положительного предсказания). Зафиксируем $\varepsilon > 0$ – точность предсказания.

Определение 6. Объект o назовем - *важным*, если суммарная вероятность появления таких ВКФ-гипотез $\langle A, B\rangle$, что $B \in H_\kappa^{\downarrow}(o)$ будет больше ε. Семейство ВКФ-гипотез назовем *-сетью*, если для каждого ε-важного объекта найдется хотя бы одна ВКФ-гипотеза из этого семейства, которая предскажет этот объект положительно.

Теорема 6. *Для n признаков и любых $\varepsilon > 0$ и $1 > \delta > 0$ достаточно породить*

$$N \geq \frac{2(n+1) - 2\log_2(\delta)}{\varepsilon}$$

ВКФ-гипотез, чтобы с вероятностью $> 1 - \delta$ все - важные объекты могли быть предсказаны положительно.

3. Вопросы программной реализации

Описанные выше вероятностные когнитивные процедуры, основанные на операции сходства, были запрограммированы автором в единой программной системе, получившей название ВКФ-система:

- Программа реализована как консольное приложение. Она была создана в среде Code::Blocks (version 13.12) с использованием библиотеки boost (version 1_56_0). Компилятор C++ - GNU C++ toolset (version 4.9.1).
- Примеры (обучающие, контр- и представленные для предсказания целевого свойства) представляются объектами класса *boost::dynamic_bitset*. Они сохраняются в контейнерах типа *std::vector* и *std::list* стандартной библиотеки C++.
- Программа использует классы *boost::random* для датчиков случайных чисел. Это нужно для спаривающей цепи Маркова (Алгоритм 3).
- Для реализации многопоточности используются классы *boost::thread*.
- Программа платформенно независима: она собиралась и запускалась под Windows и под Linux.

Прокомментируем некоторые достоинства ВКФ-системы по сравнению с классическим ДСМ-подходом:

- Так как каждая ВКФ-гипотеза порождается независимым запуском цепи Маркова, то ВКФ-программа использует несколько потоков для вычисления индуктивного обобщения. Для ДСМ-системы подобное распараллеливание индукции невозможно.
- ВКФ-система вычисляет процедуру абдуктивного уточнения и принятия ВКФ-гипотез тоже в несколько потоков. В ДСМ-системе распараллеливание шага абдукции возможно, но пока не реализовано.
- Предсказание свойств по аналогии осуществляется в один поток, так как вычислительная сложность этого шага мала в сравнении с шагом индукции.
- На ЦПУ с четырьмя потоками (i5-3210M) максимальная нагрузка процессора при вычислении в 4 потока достигает 90 %. Для существующей параллельной версии ДСМ-системы она не превосходит 50 %.

Дальнейшим программным улучшением следует считать реализацию ленивых вычисле-

ний. В настоящее время ВКФ-кандидаты находятся согласно Алгоритму 3 с использованием операций *CbODown* и *CbOUp*. Согласно определению 2 $CbODown(\langle A, B\rangle, o) = \langle (A \cup \{o\})'', (A \cup \{o\})'\rangle$.

Если вычисление пересечения $(A \cup \{o\})' = B \cap o'$ фрагмента текущего ВКФ-кандидата с фрагментом выбранного объекта *o* соответствует побитовому умножению соответствующих строк, то операция $(A \cup \{o\})'' = (B \cap o')'$ формирования нового списка родителей может потребовать побитово перемножить с полученным ранее пересечением почти все объекты, чтобы проверить, обладает ли еще какой-нибудь объект полученным пересечением.

Для улучшения ситуации предлагается (лениво) откладывать вычисления второй производной, пока последовательный выбор нескольких объектов для *CbODown* не сменится выбором признака с переходом к операции *CbOUp*. Аналогично, операция *CbOUp* имеет в своем составе потребляющую много времени компоненту $(B \cup \{f\})'' = (A \cap f')'$. Здесь тоже можно лениво откладывать вычисления этой части до тех пор, пока выбор нескольких признаков для *CbOUp* не сменится выбором объекта с переходом к операции *CbODown*.

Возникает вопрос о степени экономии, достигаемой такой процедурой. Мы предложим анализ этой проблемы с помощью техники рекуррентных событий [14].

Теорема 7. *На формальном контексте с k примерами, описываемыми n признаками, выигрыш от использования ленивых вычислений составляет* $\frac{n}{k} + \frac{k}{n} \geq 2$ *раз.*

Во время проведения экспериментов с ВКФ-системой был обнаружен феномен очень быстрого нахождения очередного ВКФ-кандидата. Хотя мы не смогли получить оценку в общем виде, для случая Булеана имеются результаты о среднем времени склеивания и сильной концентрации времени склеивания около своего среднего.

До конца этого раздела мы ограничимся случаем Булеана. Пусть $O = \{o_1, o_2, \dots, o_n\}$ будет множеством объектов, каждый из которых описывается признаками из списка $F = \{f_1, f_2, \dots, f_n\}$, и $o_i I f_j \Leftrightarrow i \neq j$ (Табл.1).

Табл. 1

O\F	f_1	f_2	...	f_n
o_1	0	1	...	1
o_2	1	0	...	1
⋮	⋮	⋮	⋱	⋮
o_n	1	1	...	0

Ясно, что

$$o_{j_1} \cap o_{j_2} \cap \dots \cap o_{j_k} = F \setminus \left\{f_{j_1}, f_{j_2}, \dots, f_{j_k}\right\},$$

так как добавление в сходство примера o_k с номером *k* удаляет из фрагмента признак f_k с тем же самым номером *k*. Очевидно, что таким образом может быть получено любое подмножество - элементного множества *F*. В дальнейшем нам будет полезен явный вид операций «замыкай-по-одному» в решетке-Булеане

$$o_i I f_j \Leftrightarrow i \neq j,$$

для множества объектов $O = \{o_1, o_2, \dots, o_n\}$, каждый из которых описывается признаками из списка $F = \{f_1, f_2, \dots, f_n\}$. Ясно, что в этом случае

$$CbODown(\langle A, B\rangle, o_k) = \\ = \begin{cases} \langle A \cup \{o_k\}, B \setminus \{f_k\}\rangle, & \text{если } o_k \notin A \\ \langle A, B\rangle & \text{иначе,} \end{cases}$$

так как добавление в сходство примера o_k с номером *k* удаляет из фрагмента признак f_k с тем же самым номером *k*. Аналогично

$$CbOUp(\langle A, B\rangle, f_k) = \\ = \begin{cases} \langle A \setminus \{f_k\}, B \cup \{f_k\}\rangle, & \text{если } f_k \notin B \\ \langle A, B\rangle & \text{иначе,} \end{cases}$$

так как добавление в сходство признака f_k с номером *k* удаляет из родителей сходства объект o_k с тем же самым номером *k*.

Определение 6.*Расстояние* $\rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle)$ между кандидатами $\langle A_1, B_1\rangle$ и $\langle A_2, B_2\rangle$**.** определяется как число позиций, в которых отличаются битовые строки B_1 и B_2. Другими словами, расстояние равно минимальному количеству ребер между соответствующими вершинами гиперкуба. Это расстояние на гиперкубе называется *метрикой Хэмминга*. Оказывается, что после применения операций «замыкай-по-одному» в случае Булеана, это расстояние не увеличивается.

Теорема 8. *Среднее время склеивания для n-мерного гиперкуба*

$$\boldsymbol{E}[\textstyle\sum_{j=1}^{n} T_j] = \sum_{j=1}^{n} \frac{n}{j} \approx n \cdot \ln(n) + n \cdot \gamma + \frac{1}{2}.$$

Теорема 9. $\boldsymbol{P}[\sum_{j=1}^{n} T_j \geq (1+\varepsilon) \cdot n \cdot \ln(n)] \to 0$ *при* $n \to \infty$ *для любого* $\varepsilon > 0$.

4. Направление дальнейших исследований

Теперь сформулируем проблемы, требующие дальнейших исследований:

- Исследовать вопрос о времени перемешивания для немонотонной цепи Маркова. Следует отметить, что в частном случае Булеана, подобный результат известен [18].
- Получить оценку среднего времени склеивания в общем случае. Полезно указать, что метрика Хэмминга между верхним и нижним ВКФ-кандидатом, не является функцией Ляпунова (может возрастать) в спаривающей цепи Маркова. Соответствующий пример есть у автора в статье [5].

Заключение

Статья описывает современное состояния дел с исследованием ВКФ-метода интеллектуального анализа данных. Были сформулированы основные известные результаты о вероятностных вариантах когнитивных процедур, основанных на операции сходства. В разделе 4 перечислены открытые проблемы ВКФ-метода, к исследованию которых призывается читатель.

Автор благодарит проф. В.К. Финна за внимание к работе, проф. Е.М Бениаминова за полезные обсуждения и своих коллег по Лаборатории 35 ФИЦ ИУ РАН за поддержку и полезные дискуссии.

Литература

1. Аншаков О.М., Скворцов Д.П., Финн В.К. Логические средства экспертных систем типа ДСМ // Семиотика и информатика. – Вып. 28. – 1986. – С. 65–102.
2. Аншаков О.М., Скворцов Д.П., Финн В.К. О дедуктивной имитации некоторых вариантов ДСМ-метода автоматического порождения гипотез // Семиотика и информатика. - Вып. 33. – 1993. – С. 164–233.
3. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). – М.: Наука. – 1974. – 416 с.
4. Виноградов Д.В. Формализация правдоподобных рассуждений в логике предикатов // Научная и техническая информация, Сер. 2. – 2000. – № 11. – С. 17–20.
5. Виноградов Д.В. Вероятностное порождения гипотез в ДСМ-методе с помощью простейших цепей Маркова // Научная и техническая информация. Сер. 2. – 2012. – № 9. – С. 20–27.
6. Виноградов Д.В. Вероятность порождения случайного ДСМ-сходства при наличии контр-примеров // Научная и техническая информация. Сер. 2. – 2015. – № 3. – С. 1–5.
7. Гусакова С.М., Финн В.К. Сходства и правдоподобный вывод // Известия АН СССР, Сер. «Техническая кибернетика». – 1987. – № 5. – С. 42–63.
8. ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания (ред.: Финн В.К., Аншаков О.М.) – М.: URSS.– 2009. – 432 с.
9. Кузнецов С.О. Интерпретация на графах и сложностные характеристики задач поиска закономерностей определенного вида // Научная и техническая информация. Сер. 2. – 1989. – № 1. – С. 23–28.
10. Кузнецов С.О. Быстрый алгоритм построения всех пересечений объектов из нижней полурешетки // Научная и техническая информация. Сер. 2. – 1993. – № 1. – С. 17–20.
11. Милль Дж.Ст. Система логики силлогистической и индуктивной: Изложение принципов доказательства в связи с методами научного исследования. Пер. с англ. Изд. 5. – М.: URSS.–2011. – 832 с.
12. Пирс Ч.С. Рассуждение и логика вещей: Лекции для Кэмбриджских конференций 1898 года. Пер. с англ. – М.: РГГУ– 2005. – 371 с.
13. Скворцов Д.П. О некоторых способах построения логических языков с кванторами по кортежам // Семиотика и информатика. – Вып. 20. – 1983. – С. 102–126.
14. Феллер В. Введение в теорию вероятностей и ее приложения. В 2–х томах. Т. 1: Пер. с англ. – М.: Мир.– 1984. – 528 с.
15. Финн В.К. Базы данных с неполной информацией и новый метод автоматического порождения гипотез // В кн.: Диалоговые и фактографические системы информационного обеспечения. – М.– 1981. – С. 153–156.
16. Финн В.К. Синтез познавательных процедур и проблема индукции // Научная и техническая информация. Сер. 2. – 1999. – № 1–2. – С. 8–45.
17. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта. – 2004. – № 3. – С. 3–18.
18. Diaconis, Persi, Group representations in probability and statistics. IMS Lecture Notes – Monograph Series Vol. 11. – Hayward (CA): Institute of Mathematical Statistics, 1988. – 198 pp.
19. Finn, V.K., Mikheyenkova, M.A. Plausible Reasoning for the Problems of Cognitive Sociology // Logic and Logical Philosophy, Vol. 20. – 2011. – p. 111-137.
20. Ganter, Bernard and Wille, Rudolf, Formal Concept Analysis. Transl. from German. – Berlin: Springer–Verlag, 1999. – 284 pp.

Виноградов Дмитрий Вячеславович. Старший научный сотрудник ФИЦ ИА РАН. Окончил МГУ им. М.В. Ломоносова в 1986 году. Кандидат физико-математических наук. Количество печатных работ: 25. Область научных интересов: искусственный интеллект, математическая логика, дискретная математика. Email: KRRGuest@yandex.ru

VKF-method of intelligent data analysis: current state of the art and open problems
D.V. Vinogradov

Abstract. The paper describes current state of the art for VKF-method of intelligent data analysis. This method combines three cognitive procedures (induction, abduction, and analogy) based on probabilistic algorithm for similarity calculation. We demonstrate main results and formulate open problems to investigate them.
Keywords: similarity, Markov chain, VKF-candidate, counter-example, prediction by analogy.

References

1. Anshakov, O.M., Skvortsov, D.P., Finn, V.K. Logicheskie sredstva expertnyh system tipa JSM // Semiotica i informatika. – Issue 28. – 1986. – p. 65–102.
2. Anshakov, O.M., Skvortsov, D.P., Finn, V.K. O deduktivnoj imitacii nekotoryh variantov JSM-metoda avtomaticheskogo porozhdenija gipotez // Semiotica i informatica. - Issue 33. – 1993. – p. 164–233.
3. Vapnik, V.N., Chervonenkis, A.Y. Teorija raspoznavania obrazov. – M.: Nauka. – 1974. – 416 pp.
4. Vinogradov, D.V. Formalizing plausible arguments in predicate logic // Autom. Doc. Math. Linguist Vol. 34 – 2000. – № 6. – p. 6–10.
5. Vinogradov, D.V. Random generation of hypotheses in the JSM method using simple Markov chains // Autom. Doc. Math. Linguist Vol. 46 – 2012. – № 5. – p. 221–228.
6. Vinogradov, D.V. The probability of encountering an accidental DSM similarity in the presence of counter examples // Autom. Doc. Math. Linguist Vol. 49 – 2015. – № 2. – p. 43–46.
7. Gusakova, S.M., Finn, V.K. Skhodstvo i pravdopodobnyj vyvod // Izvestija AN SSSR, Ser. «Tehnicheskaja kibernetika». – 1987. – № 5. – p. 42–63.
8. JSM-metod avtomaticheskogo porozhdenija hypotez: Logigeskie i epistemologicheskie osnovanija. (Eds.: Finn, V.K., Anshakov, O.M.) – M.: URSS, 2009. – 432 pp.
9. Kuznetsov, S.O. Interpretacija na grafah i slozhnostnye harakteristiki zadach poiska zakonomernostej opredelennogo tipa // Nauchnaja i tehnicheskaja informacija, Ser. 2. – 1989. – № 1. – p. 23–28.
10. Kuznetsov, S.O. Bystryj algoritm postroenija vseh peresechenij objektov iz nizhnej polureshetki // Nauchnaja i tehnicheskaja informacija, Ser. 2. – 1993. – № 1. – p. 17–20.
11. Mill, J.S. A System of Logic, Ratiocinative and Inductive – Honolulu, 2002. – 644 pp.
12. Peirce, C.S. Reasoning and the Logic of Things – Boston: Harvard University Press, 1993. – 312 pp.
13. Skvortsov, D.P. O nekotoryh sposobah postroenija logicheskih jazykov s kvantorami po kortezham // Semiotica i informatica. – Issue 20. – 1983. – p. 102–126.
14. Feller, W. An Introduction to Probability Theory and Its Applications. Vol. 1, 3rd Ed. – NY: Wiley, 2008. – 510 pp.
15. Finn, V.K. Bazy dannyh s nepolnoj informaciej i novyj metod avtomaticheskogo porozhdenija hypotez // In: Dialogovye i faktograficheskie sistemy informacionnogo obespechenija. – M., 1981. – p. 153–156.
16. Finn, V.K. The synthesis of cognitive procedures and the problem of induction // Autom. Doc. Math. Linguist Vol. 43 – 2009. – № 3. – p. 149–195.
17. Finn, V.K. Ob intellektualnom analize dannyh // Novosti iskusstvennogo intellekta. – 2004. – № 3. – p. 3–18.
18. Diaconis, Persi, Group representations in probability and statistics. IMS Lecture Notes – Monograph Series Vol. 11. – Hayward (CA): Institute of Mathematical Statistics, 1988. – 198 pp.
19. Finn, V.K., Mikheyenkova, M.A. Plausible Reasoning for the Problems of Cognitive Sociology // Logic and Logical Philosophy, Vol. 20. – 2011. – p. 111-137.
20. Ganter, Bernard and Wille, Rudolf, Formal Concept Analysis. Transl. from German. – Berlin: Springer–Verlag, 1999. – 284 pp.

Vinogradov Dmitry V. Senior research scientist at Federal Research Center for Computer Science and Control, Russian Academy of Science, 40 Vavilova Street, Moscow 119333, Russia. PhD on Theoretical Foundations of Informatics. Published 25 papers indexed by Russian Index for Scientific Citations. Research areas: AI, mathematical logic, discrete mathematics