А.А. Онищенко, С.И. Гуров

КЛАССИФИКАЦИЯ НА ОСНОВЕ АФП И БИКЛАСТЕРИЗАЦИИ: ВОЗМОЖНОСТИ ПОДХОДА*

1. Введение

Неформальная постановка рассматриваемых ниже *задач классификации по прецедентам* заключается в следующем. Имеется множество объектов X, разбитое на два класса X^+ (*положительный*) и X^- (*отрицательный*) относительно обладания/необладания объектами некоторого *целевого признака*. Элементы данных классов называются, соответственно, *положительными* или *отрицательными примерами*. Информация о таком разбиении содержится только в указании о принадлежности к данным классам элементов конечной *обучающей последовательности* (или *выборки*) из X, элементы которой называют *прецедентами*. Все объекты имеют описание на определённом формальном языке, указывающем степень обладания объектами конечным числом некоторых признаков. Указанное описание прецедентов мы представляем в виде *объектно-признаковой* $(0,1)$*-матрицы*, в которой объектам соответствуют строки, признакам — столбцы, а элементы матрицы кодируют наличие/отсутствие признаков у объектов. По данному описанию прецедентов необходимо сформулировать *решающее правило* или *классификатор*, который по описанию нового объекта из X указывал бы имя класса, его содержащего. При этом решающее правило должно обладать свойством оптимальности по отношению к некоторому функционалу, определяющему качество классификации.

Значительная группа современных методов поиска зависимостей и анализа данных базируется на теории решёток замкнутых множеств (решёток формальных понятий или решёток Галуа). К ним относится анализ формальных понятий (АФП, *англ. Formal Concept Analysis*) [1,2]. Важной стороной методов классификаций на основе АФП является отсутствие тех или иных допущений относительно появления объектов в данной задаче (такие допущения, обычно не проверяемые на практике, могут отражаться в построенных распознающих алгоритмах в виде тех или иных условий, в действительности не присущих рассматриваемой задаче). Развитием метода АФП является подход, называемый бикластеризацией, и представляющий собой совокупность моделей и методов, альтернативных

* Работа выполнена при частичной финансовой поддержке РФФИ, проект № 10-01-00131-а и ЗАО «Интел».

классическим подходам к кластеризации и опирающихся на идею сохранения объектно-признакового описания сходства кластеров.

С помощью АФП и методов на его основе решают задачи анализа данных и обработки знаний, в частности, и прикладные задачи классификации по положительным и отрицательным примерам [12]. Целью данной работы является исследование эффективности таких подходов. Все описанные ниже алгоритмы реализованы в среде MatLab.

2. АФП и решётка формальных понятий

Анализ формальных понятий — прикладная область теории решёток. Пусть G и M — непустые множества, называемые соответственно *множествами объектов* и *признаков*, а I — соответствие между G и M, для которого принята инфиксная форма записи, интерпретирующаяся следующим образом: gIm означает, что объект $g \in G$ обладает признаком $m \in M$.

Определение 1. Тройка $K = (G, M, I)$ называется *формальным контекстом*.

Понятно, что в конечном случае контекст может быть задан виде объектно-признаковой $(0,1)$-матрицы.

Для произвольных $A \subseteq G$ и $B \subseteq M$ вводятся отображения $\varphi: 2^G \to 2^M$ и $\psi: 2^M \to 2^G$ такие, что

$$A\varphi = \{m \subseteq M | \forall\, g \in A\ (gIm), \quad B\psi = \{g \subseteq G | \forall\, m \in B\ (gIm).$$

(если φ — отображение P на Q, то образ $q \in Q$ элемента $p \in P$ мы обозначаем $p\varphi$, а образ подмножества $A \subseteq P$ — $A\varphi$). Нетрудно видеть, что пара отображений (φ, ψ) является соответствием Галуа между ч.у. множествами 2^G и 2^M, упорядоченными по включению [4]. В соответствии с традицией АФП, отображения φ и ψ обозначаются $(\cdot)'$, так что приведённые выше множества $A\varphi$ и $B\psi$ записываются как A' и B' соответственно. Двойное применение операции $(\cdot)'$ является, очевидно, оператором замыкания на объединении множеств 2^G и 2^M.

Определение 2. Пусть дан контекст K. Пара подмножеств (A, B), где $A \subseteq G$, а $B \subseteq M$, и таких, что $A' = B$ и $B' = A$, называется *формальным понятием* данного контекста с *формальным объёмом* A и *формальным содержанием* B.

Если контекст K представлен в виде $(0,1)$ -матрицы, то формальному понятию соответствует максимальная её подматрица, заполненная единицами.

Очевидно, что объём и содержание произвольного формального

понятия являются замкнутыми множествами. Множество всех формальных понятий $\{(A, B)\}$ данного контекста K образует полную решётку $\mathfrak{B}(K)$ относительно операций $\vee$ (объединение) и $\wedge$ (пересечение):

$$(A_1, B_1) \vee (A_2, B_2) \triangleq ((B_1 \cap B_2)', B_1 \cap B_2),$$

$$(A_1, B_1) \wedge (A_2, B_2) \triangleq (A_1 \cap A_2, (A_1 \cap A_2)'),$$

называемую *решёткой формальных понятий* или *решёткой Галуа*.

3. Основные понятия классификации на основе АФП

Пусть M — множество признаков, которые назовем *структурными*, а $w \notin M$ — некоторый *целевой* признак (свойство).

Данные для обучения представляются множествами положительных, отрицательных и недоопределённых примеров (символически $(+)$ -, $(-)$ - и (τ) -примеры соответственно). Положительные (отрицательные) примеры суть объекты, про которые известно, что они обладают (соответственно, не обладают) свойством w. Для недоопределённых примеров неизвестно значение предиката обладания свойством w, и цель классификации состоит в определении значения этого предиката.

В терминах АФП входные данные могут быть описаны с помощью трех контекстов по отношению к данному свойству w: положительного $K_+ = (G_+, M, I_+)$, отрицательного $K_- = (G_-, M, I_-)$ и недоопределённого $K_\tau = (G_\tau, M, I_\tau)$. Здесь M — множество структурных признаков, G_+, G_- и G_τ — совокупности соответственно положительных, отрицательных и недоопределённых примеров, а $I_\epsilon \subseteq G_\epsilon \times M$, где $\epsilon \in \{+, -, \tau\} \triangleq E$ — соответствия, определяющие структурные признаки $(+)$-, $(-)$- и (τ)-примеров. Операторы Галуа в этих контекстах обозначаются, соответственно, верхними индексами из E, например, A^+, A^-, A^τ. Для краткости далее (кроме случая $G, M \subseteq \mathbb{N}$) будем писать g', g'', g^-, g^τ, m' и т. д., вместо $\{g\}'$, $\{g\}''$, $\{g\}^-$, $\{g\}^\tau$, $\{m\}'$ соответственно.

Теперь могут быть определены положительная и отрицательная гипотезы в пользу положительной и отрицательной классификации соответственно по отношению к некоторому свойству [6].

Пусть задано свойство w и множество примеров, для которых определены контексты K_+, K_- и K_τ. Последующие определения даются по отношению к выбранному свойству w.

Определение 1. Формальное понятие положительного контекста называется *положительным*.

Если (A, B) — положительное понятие, то множество A называется

его *положительным формальным объёмом*, а множество B — *положительным формальным содержанием*.

Аналогично определяются отрицательные и недоопределённые понятия, формальные объём и содержание для контекстов K_- и K_τ.

Определение 2. Положительное формальное содержание B положительного понятия (A, B) называется

- *положительной* или *$(+)$-предгипотезой*, если оно не является формальным содержанием ни одного отрицательного понятия;
- *положительной* или *(минимальной) $(+)$-гипотезой*, если оно не является подмножеством содержания g^- для некоторого элементарного понятия (g, g^-) для отрицательного примера g, и *$(+)$-фальсифицированной гипотезой* или *фальсифицированным $(+)$-обобщением*, иначе.

Отрицательные (или $(-)$-) *предгипотезы, гипотезы* и *фальсифицированные гипотезы* определяются аналогично.

Из определения непосредственно следует, что гипотеза является также и предгипотезой. Гипотезы используются для классификации недоопределённых примеров из множества G_τ.

Определение 3. Если формальное содержание g^τ недоопределённого примера $g \in G_\tau$ содержит положительную (отрицательную) гипотезу, то говорят, что последняя является *гипотезой в пользу положительной (отрицательной) классификации* g соответственно.

АФП ориентирован на анализ качественной информации. Для получения количественных признаков используется процедура шкалирования. С этой целью, кроме двузначных контекстов, в АФП используются и *многозначные*, имеющие вид (G, M, W, I), где G, M, W — множества объектов, признаков и значений признаков, соответственно, а I — тернарное отношение $I \subseteq G \times M \times W$, задающее значение $w \in W$ признака $m \in M$ объекта $g \in G$, причем связанное с I отображение $G \times M \to W$ функционально.

Шкалирование есть представление многозначных контекстов двузначными. Шкалой для признака m многозначного контекста называется контекст вида (G_m, M_m, I_m) с $m(G) \subseteq G_m$. Шкалу составляют объекты (*значения шкалы*) и *признаки*. Примерами шкал являются номинальная, порядковая, межпорядковая, дихотомичная (булева), контрноминальная и т. д. шкалы [2].

Простейшая модель обучения в терминах АФП основана на общем принципе: для заданных положительных и отрицательных примеров

«целевого понятия» необходимо построить «обобщение» положительных понятий, которое не покрывало бы отрицательных. Если имеется гипотеза в пользу положительной (отрицательной) классификации и нет гипотез в пользу отрицательной (положительной) классификации, то $g \in G_\tau$ классифицируется положительно (отрицательно). Отказ от классификации происходит, если формальное содержание g^τ либо не включает в качестве подмножеств ни положительных, ни отрицательных гипотез (недостаток данных), либо включает в себя как положительные, так и отрицательные гипотезы (противоречивые данные).

4. Алгоритмы и результаты классификации «прямым» методом АФП

По описанному методу АФП был составлен алгоритм решения задач классификации. В нём предусматривалось линейное шкалирование признаков — получение вместо одного $n \times intervals$ признаков (где $intervals$ — задаваемый параметр алгоритма). Алгоритм был применён для распознавания заболевания печени (Liver Disorders) по данным анализов. Эта и некоторые из встречающихся далее в п. 6 задач взяты из банка UCI Machine Learning Repository [1]. В рассматриваемой задаче объектами являются совокупности данных шести анализов исследуемых пациентов. Имеется обучающая выборка из 345 прецедентов, разделенная на положительный и отрицательный классы относительно целевого признака «наличие заболевания печени». Сложность задачи состоит в том, что отклонение показателей анализов от нормальных значений может быть вызвано не только заболеваниями печени, но и другими причинами.

Реализованный алгоритм был протестирован на данной задаче методом скользящего контроля. Результат состоял в том, что алгоритм практически всегда отказывается от классификации по недостатку информации за исключением всего лишь пяти случаев, которые классифицируются правильно.

Ясно, что данный алгоритм непригоден для решения поставленной задачи. С целью его улучшения была проведена следующая модификация, касающаяся порождения гипотез и способа классификации.

1. Модификация гипотез — в гипотезу добавлялись признаки, которыми обладали “почти” все объекты заданного класса. При этом контролировалось, чтобы доля объектов класса, отклоняющихся от гипотезы, не превышала некоторого значения P (новый параметр программы) и, понятно, уже не гарантируется, что гипотеза не входит ни в

[1] http://archive.ics.uci.edu/ml

одно описание объекта из другого класса.

Заметим, что в данном направлении ведутся интенсивные исследования [2,9,10].

2. Введение метрики между объектами и модификация классификации — чем в большем числе координат объекты различаются, тем больше между ними «расстояние». Таким образом, вычислялось расстояние между гипотезами положительного и отрицательного классов и классифицируемым объектом. Полученные расстояния делились на количество единиц в соответствующих гипотезах. Объекту присваивался тот класс, с которым у него было меньше различий.

3. Введение весов признаков — признак имеет тем больший вес, чем больше единиц содержит соответствующий ему столбец.

Полученный модифицированный алгоритм был применён к рассматриваемой задаче Liver Disorders. В результате при значениях параметров $intervals = 10$ и $P = 0.2$ алгоритм отказывается классифицировать 266 (77%) объектов; из оставшихся 79 объектов 54 классифицируются верно и 25 — ошибочно, т. е. процент ошибок $\approx 32\%$.

5. Бикластеризация

Развитием подхода к классификации на основе АФП является метод *бикластеризации* [7, 8, 9].

Используя методы АФП, для любых объектно-признаковых данных можно построить иерархическую структуру формальных понятий (бикластеров), позволяющую отразить их таксономические свойства в удобном для аналитика виде. Основным недостатком решёток понятий является их большой размер: для объектно-признаковой таблицы размером $m \times n$ число таких бикластеров в худшем случае равно 2^n. Идея рассматриваемого подхода состоит в ослаблении требований к формальным понятиям, что даёт возможность не только сократить число порождаемых бикластеров, но и устранить влияния шума (вариаций признаков) на результаты.

Определение 1. Для формального контекста $K = (G, M, I)$ *объектным понятием* называется формальное понятие вида (g'', g'), где $g \in G$, а *признаковым понятием* — формальное понятие вида (m', m''), где $m \in M$.

Определение 2. Для формального контекста $K = (G, M, I)$ и любой пары объектных и признаковых понятий (g'', g') и (m', m''), связанных отношением вложения $(g'', g') \leq (m', m'')$, назовем *бикластером* пару вида (m', g').

Бикластер есть подматрица объектно-признаковой $(0,1)$-матрицы, такая, что её строки проявляют «сходство друг с другом» на столбцах и наоборот.

Определение 3. *Плотностью* бикластера (A, B) формального контекста $K = (G, M, I)$ называется величина

$$\rho(A, B) = |I \cap \{A \times B\}|/(|A| \cdot |B|).$$

Очевидно, что $0 \leq \rho(A, B) \leq 1$, а если (A, B) — формальное понятие, то $\rho(A, B) = 1$.

Зададимся некоторым числом $\rho_{min} \in [0, 1]$ и будем называть бикластер (A, B) *плотным*, если $\rho(A, B) \geq \rho_{min}$, и определим на бикластерах отношение вложения $\sqsubseteq$:

$$(X_1, Y_1) \sqsubseteq (X_2, Y_2) \triangleq (X_1 \subseteq X_2) \text{ и } (Y_1 \subseteq Y_2).$$

Оказывается, что при $\rho_{min} = 0$ для любого формального понятия некоторого контекста K существует бикластер, в который оно вкладывается. С другой стороны, при достаточно больших значениях ρ_{min} не все формальные понятия могут оказаться вложенными в некоторый бикластер, построенный по данному формальному контексту. Существует быстрый (со сложностью, не более, чем $O(|G| \cdot |M|)$) алгоритм поиска бикластеров [9].

Пусть $K = (G, M, I)$ — формальный контекст, (A, B) — некоторое формальное понятие K, тогда *индекс устойчивости* σ понятия (A, B) определяется выражением $\sigma(A, B) = |\mathcal{C}(A, B)|/2^{|A|}$, где $\mathcal{C}(A, B)$ — объединение подмножеств $C \subseteq A$, таких, что $C = B'$. Очевидно, что $0 \leq \sigma(A, B) \leq 1$. Если выбрано значение $\sigma_{min} \in [0, 1]$, то формальное понятие (A, B) назовём *устойчивым*, если $\sigma(A, B) \geq \sigma_{min}$.

Бикластеры, а также плотные и устойчивые формальные понятия используют для формирования гипотез при решении задач кластеризации [10].

6. Применение метода бикластеризации для классификации

С точки зрения практического применения при классификации бикластеризация может быть использована как обработка данных после шкалирования: для последующего порождения гипотез отбираются лишь «значимые» объекты, т. е. те, чья плотность превышает порог ρ_{min}. Такой подход позволяет избегать использования шумовых эффектов при порождении гипотез.

Отличие от вышеописанного алгоритма состоит в том, что теперь

гипотезы порождаются с использованием отобранных бикластеризацией «хороших объектов». В описанном методе присутствуют два настраиваемых параметра: ρ_{min} — порог плотности бикластера и P — доля объектов класса, отклоняющихся от классической гипотезы (используется в процедуре порождения гипотез). Подбор этих параметров нетривиален.

Если значение параметра ρ_{min} будет занижено, то в порождении гипотез будут участвовать шумовые признаки и выбросы. Если же его значение будет завышено, будут завышены и требования к гипотезе. Иными словами, порожденная гипотеза будет указывать на то, что все объекты класса обладают значительным набором признаков, но вероятность того, что у классифицируемого объекта будут присутствовать все эти признаки, мала. Возможно, было бы полезным использовать интервал для значений параметра ρ_{min}, чтобы отбирать основные, а не граничные объекты. Этот метод был протестирован, однако существенных улучшений качества классификации не дал, что будет объяснено позже спецификой именно этой конкретной задачи.

Когда значение параметра P близко к нулю, гипотезы порождаются согласно классическим представлениям АФП: в них входят только те признаки, которыми обладают все объекты данного класса. Проблема состоит в том, что если класс объединяет большое число объектов, то признаков, которыми обладают все объекты, будет очень мало, и гипотеза теряет свою «представительность» для данного класса. Более того, велика вероятность гипотезы одного класса совпасть с гипотезой другого класса. Если же значение параметра взять слишком большим, то гипотеза будет требовать от контрольного объекта обладать большим количеством признаков, что опять же может быть слишком жёстким условием. В некотором смысле, это — известный в распознавании образов эффект *переобучения*.

По вышеописанному алгоритму была составлена компьютерная программа. Результаты тестирования на различных задачах методом скользящего контроля представлены в приведённой ниже таблице.

В заголовке таблицы n — число признаков, l — число объектов (длина обучающей выборки). Далее в столбцах приведены результаты решения задачи классификации при оптимизации параметров алгоритма (порога ρ и доли P) по двум критериям: % ошибок классификации err и числу отклассифицированных объектов l_c ($l - l_c =$ число отказов от классификации). Локальная оптимизация параметров алгоритма производилась методом Гаусса-Зейделя, а их оптимальные значения ρ_{min} и P^* даны вместе с err.

Данные первых четырёх задач имеются в репозитории UCI Machine Learning (ссылка дана выше). Задача № 5 (Two norm) на разделение двух нормальных 20-мерных распределений взята с сайта Университета в Торонто[2]; алгоритм классификации CART (см. *Breiman L., Friedman J. H., Olshen R. A. and Stone C. J.* Classification and Regression Trees. Wadsworth International Group: Belmont, California) при длине обучающей выборки в 300 прецедентов показал на ней 22.1% ошибок, что почти в 10 раз больше теоретического минимума для идеального классификатора – дискриминантной функции Фишера. Задачи № 6 Lung cancer (Рак легких), № 7 Cirrhosis (Цирроз печени) и № 8 Cloud-seeding (Наблюдение облачности) взяты с сайта системы StatLib статистического программного обеспечения[3].

№	Задача	n	l	l_c	err ρ_{min}; P^*	l_c	err ρ_{min}; P^*
1.	Liver Disorders	6	345	22	13.6% 0.30; 0.01	79	29.1% 0.30; 0.20
2.	Glass identification	9	146	28	10.00% 0.15; 0.05	59	16.90% 0.30; 0.20
3.	Wine	13	130	76	02.00% 0.25; 0.05	85	08.20% 0.30; 0.20
4.	Wine quality	11	130	83	08.40% 0.25; 0.05	141	13.50% 0.30; 0.20
5.	Two norm	20	354	206	12.10% 0.15; 0.15	233	15.20% 0.30; 0.20
6.	Lung cancer	8	137	18	05.50% 0.01; 0.01	98	25.50% 0.05; 0.05
7.	Cirrhosis	19	276	33	21.00% 0.05; 0.05	83	37.79% 0.30; 0.20
8.	Cloud-seeding	5	108	7	28.00% 0.15; 0.05	20	30.00% 0.15; 0.15

6. Анализ результатов и выводы

С одной стороны, относительно свойств АФП в литературе можно найти утверждения, что он может быть применён для решения некоторых задач распознавания образов, так как представляет собой удобное средство для формализации символьных моделей машинного обучения. Также

[2] http://www.cs.toronto.edu/ delve/data/twonorm/desc.html

[3] http://lib.stat.cmu.edu/datasets, страницы /veteran, /pbc и /cloud соответственно.

утверждается, что данный метод «*нашел успешное широкое применение в информатике, в частности, при решении задач классификации по положительным и отрицательным примерам, медицинской диагностики* (и др.)» [12].

С другой — для полученных нами данных справедливы следующие наблюдения.

1. Применение метода бикластеризации с оптимизацией параметров позволило улучшить качество классификации относительно результатов модернизированного алгоритма АФП весьма незначительно (в задаче № 1 — лишь на 3%).
2. Во всех случаях наблюдалась достаточно высокая и, как правило, неприемлемая на практике доля отказов от классификации.
3. Во всех случаях наблюдался достаточно высокий и, как правило, неприемлемый на практике процент ошибок.
4. При попытках подстройки параметров алгоритма с целью уменьшения доли отказов обычно наблюдался сильный — в разы — рост количества ошибок, хотя встречаются и случаи его незначительного роста (задача № 8); при этом число прецедентов, поддающихся классификации, могло вырасти весьма значительно (задача № 6).

Кроме того, анализ свойств порождаемых гипотез при разных значениях параметров и критериев оптимизации показал, что часто гипотезы разных классов вкладывались друг в друга. Геометрически это означает сильное пересечение выпуклых оболочек классов, а данный факт составляет специфику рассмотренных задач. Естественно предположить, что если бы классы менее диффундировали друг в друга, бикластеризация и классический метод АФП могли бы показать более впечатляющие результаты. Для «улучшения» расположения классов в теории распознавания образов применяют методы преобразования признакового пространства. Однако не следует забывать, что практические задачи классификации могут и не иметь удовлетворительных решений в исходных постановках. В этих случаях возможно эффективно применение методов повышения компактности данных [11].

В результате анализа применения подходов на основе АФП к решению задач классификации, можно сделать следующие выводы.

1. Без модификации и/или глубокой предобработки данных методы классификации на основе АФП могут использоваться лишь на этапах предварительной классификации.
2. Известная идея в модификации прямого подхода АФП — развитие методов порождения гипотез. Полезными здесь является учёт специфики конкретной предметной области и «подстройка» под неё гипотез и алгоритмов; при этом важную информацию

могут нести значения параметров ρ_{min}, P^*, σ_{min} и им подобных.

3. Возможна также разработка и применение более тонких правил классификации, использующих, например, веса объектов, признаков, гипотез и т.д.
4. Перспективным направлением является построение на основе АФП методов преобразования признакового пространства, в т. ч. с использованием оценок компактности данных.

Литература

1. Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. - Springer, 1999. - 314 c.
2. Кузнецов С. О. Теория решёток для интеллектуального анализа данных. [HTML] (http://vorona.hse.ru/sites/infospace/podrazd/facul/facul_bi/opm/DocLib3/ИОПF/\\book.pdf).
3. Биркгоф Г. Теория решёток. - М.: Наука, 1984. - 380 с.
4. Оре О. Теория графов. - М: Наука, 1980. - 336 с.
5. Гуров С. И. Упорядоченные множества и универсальная алгебра (вводный курс). - М.: Издат. отд. ф-та ВМиК МГУ, 2004. - 100 с.
6. Финн В. К. О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф.Бэкона-Д.С.Милля. / Семиотика и информатика. - 1983., вып. 20. - С. 35-101.
7. Mirkin B. G. Mathematical Classification and Clustering. Kluwer Academic Publishers, 1996.
8. Игнатов Д. Методы бикластеризации для анализа Интернет-данных. [HTML] (http://citforum.ru/consulting/BI/biclustering).
9. Игнатов Д. И. Модели, алгоритмы и программные средства бикластеризации на основе замкнутых множеств. Автореферат дисс. уч. степ. канд. техн. наук. Спец. 05.13.18 «Математ. моделирование, числен. методы и комплексы программ».
10. Kuznetsov S. O. On stability of a formal concept. In San Juan, E., ed.: JIM, Metz, France (2003).
11. Гуров С. И., Долотова Н. С., Фатхутдинов И. Н. «Некомпактные» задачи распознавания. Синтез схем по Э. Гильберту. / Spectral and Evolution Problems: Proceedings of the 17th Crimean Autumn Mathematical School-Symposium. - Simferopol: Crimean Scientific Center of Ukrainian National Academy of Sciences. - 2007. - V. 17. - C. 37-44.
12. Анализ формальных понятий.

    [HTML] (http://www.machinelearning.ru/wiki/index.php?title=Анализ формальных понятий).