Повышение эффективности алгоритма классификации на основе Анализа Формальных Понятий

Прокашева О. В.
prokasheva@gmail.com
МГУ имени М.В. Ломоносова

В настоящей работе исследуется метод классификации на основе построения минимальных гипотез с использованием решётки формальных понятий. Алгоритм тестируется на реальных данных с номинальными и вещественными признаками. Также сравниваются различные модификации метода для уменьшения количества отказов от классификации на основе введения метрик и процедуры голосования.

Ключевые слова: *анализ формальных понятий, классификация, машинное обучение.*

Efficiency improvement of the FCA-based classification algorythm

Prokasheva O. V
Lomonosov Moscow State University

In this paper we investigate the FCA classification algorithm on the basis of minimal hypothesis. The algorithm is tested on various benchmarks with nominal and real features. Also various modifications of the algorithm are presented and compared in terms of their efficiency. These modifications are aimed to reduce classification errors and refusals by using various metrics and voting procedures. **Keywords**: *formal concept analysis, classification, machine learning.*

Введение

Анализ формальных понятий — относительно новый метод анализа данных, относящейся к прикладной ветви алгебраической теории решёток. Данный метод нашел широкое применение в различных областях машинного обучения, таких как информационный поиск, обработка документов и текстов, построение таксономий и классификации.

Существует несколько подходов к классификации в рамках АФП [3]. В данной работе исследуется подход классификации на два класса на основе ДСМ-метода генерации гипотез [2]. В рамках данного подхода гипотезой называется некоторый набор признаков, который присутствует в описании объектов одного класса и не присутствует в описании объектов из других классов. Таким образом гипотезы способны классифицировать новые недоопределенные объекты. Согласно методу АФП гипотезы извлекаются из решёток понятий.

Ранее в работе [1] проводилась оценка эффективности метода классификации на основе генерации гипотез. Метод основан на поиске гипотез, которым удовлетворяет большинство объектов из класса, для каждого класса была построена одна гипотеза. Было показано, что алгоритм в большинстве случаев отказывается от классификации по недостатку информации.

В настоящей работе для каждого класса строится набор всевозможных минимальных гипотез. Алгоритм классификации с использованием извлеченных из обучающей выборки гипотез тестировался на реальных данных. Было установлено, что алгоритм в большом количестве случаев отказывается от классификации по противоречию, то есть гипотезы

относят объект сразу к двум классам. Поэтому было предложено применять различные модификации алгоритма для уменьшения отказов от классификации. В частности были использованы метрики и процедура голосования по большинству.

Анализ Формальных Понятий. Основные определения

Напомним некоторые основные определения Анализа Формальных понятий, согласно [2].

Определение 1. *Формальный контекст — тройка* $K = (G, M, I)$, *где* G — *множество объектов,* M — *множество признаков,* I — *соответствие между* G *и* M: gIm *означает, что объект* $g \in G$ *обладает признаками* $m \in M$

Определение 2. *Для произвольных* $A \subseteq G$ *и* $B \subseteq M$ *определены операторы Галуа:*

$$A' = \{m \in M | \forall g \in A\ (gIm)\};$$

$$B' = \{g \in G | \forall m \in B\ (gIm)\}$$

Оператор $''$ (композиция двух применений оператора $'$) — *оператор замыкания.*

Определение 3. *Пара множеств* $H = (A, B) : A \subseteq G, B \subseteq M$, $A' = B$ *и* $B' = A$, *называется формальным понятием контекста* K *с формальным объемом* A $(Extent(H) = A)$ *и формальным содержанием* B $(Intent(H) = B)$.

Определение 4. *Понятия* (A_1, B_1) *и* (A_2, B_2) *связаны отношением частичного порядка* $(A_1, B_1) \leqslant (A_2, B_2)$, *если* $A_1 \subseteq A_2$ *(что эквивалентно* $B_2 \subseteq B_1$*)*

Определение 5. *Частично упорядоченное по вложению объемов множество формальных понятий контекста* K *обозначается* $L(K)$ *и называется решеткой понятий контекста* K.

Далее рассматривается задача классификации по положительным и отрицательным примерам.

Пусть у нас имеется множество объектов G. Все объекты имеют описание на некотором формальном языке, указывающем степень обладания объектами некоторыми признаками, множество которых обозначим M. Множество G разбито на два непересекающихся класса G_+ (*положительный*) и G_- (*отрицательный*) относительно обладания их объектами некоторым *целевым признаком* $z \notin M$. Элементы данных классов, предъявленные в качестве исходных данных для решения задачи классификации, называются, соответственно, *положительными* или *отрицательными примерами.*

В терминах АФП задача классификации по положительным и отрицательным примерам формулируется следующим образом [2] :

Входные данные:

$K_+ = (G_+, M, I_+)$ – положительный контекст по отношению к целевому признаку z;
$K_- = (G_-, M, I_-)$ – отрицательный контекст по отношению к целевому признаку z;
$K_\tau = (G_\tau, M, I_\tau)$ – недоопределённый контекст по отношению к целевому признаку z;

Здесь M — множество признаков, G_+, G_- и G_τ — совокупности соответственно положительных, отрицательных и недоопределённых примеров, а $I_\varepsilon \subseteq G_\varepsilon \times M$, где $\varepsilon \in \{+ +, -, \tau\} \triangleq E$ — соответствия, определяющие их признаки.

Задача:

Для недоопределённых объектов G_τ определить неизвестное значение предиката обладания признаком z.

Определение 6. *Формальное понятие положительного контекста K_+ называется положительным. Если (A, B) — положительное понятие, то множество A называется его положительным формальным объёмом, а множество B — положительным формальным содержанием.*

Аналогично для отрицательных и недоопределённых понятий, формальных объёмов и содержаний контекстов K_- и K_τ.

Определение 7. *Положительное формальное содержание B_+ положительного понятия $(A_+, B_+) \in K_+$ называется:*

1) *положительной гипотезой, если $\underset{K_-}{\forall} (A_-, B_-)\ (B_+ \neq B_-)$, т. е. оно не является формальным содержанием ни одного отрицательного понятия;*
2) *фальсифицированной положительной гипотезой, если $\underset{K_-}{\exists} (g, g^-)\ (B_+ \subseteq g^-)$.*

Аналогично для отрицательных и фальсифицированных гипотез.

Модель классификации 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 формальное содержание g^τ содержит положительную (отрицательную) гипотезу, то говорят, что последняя является гипотезой в пользу положительной (отрицательной) классификации g соответственно.

Если формальное содержание g^τ содержит одновременно или не содержит положительную и отрицательную гипотезу, то алгоритм отказывается от классификации по противоречию (в 1 случае) и по недостатку гипотез (во 2 случае).

На практике удобно выделять не все положительные H_+(отрицательные H_-) гипотезы, а минимальные положительные (отрицательные) гипотезы, которые эвивалентны множествам всех гипотез по отношению к возможным классификациям.

Определение 8. *Положительная гипотеза h_+ есть минимальная положительная гипотеза, если ни одно подмножество $h \in h_+$ не является положительной гипотезой.*

Минимальные отрицательные гипотезы определяются аналогично.

Алгоритм классификации

В статье использован следующий алгоритм классификации по положительным и отрицательным примерам.

Алгоритм распознавания:

1) Бинаризация признаков;
2) Вычисление минимальных гипотез, соответствующих положительным и отрицательным примерам;
3) Классификация недоопределенных примеров на основе выбранных гипотез:
   (а) Если объект содержит гипотезы только из положительного (отрицательного) класса, то объект классифицируется положительно (отрицательно);
   (б) Если объект содержит гипотезы из обоих классов, алгоритм отказывается от классификации по противоречию;
   (в) Если объект не содержит никакие гипотезы из обоих классов, алгоритм отказывается от классификации по недостатку информации.

Поскольку АФП применим к данным с номинальными признаками, то в случае данных с вещественными признаками приходится применять процедуру обработки.

Задача бинаризации признаков сама по себе является отдельным полем для исследования. Применение более тонких методов обработки признакового пространства способно существенно улучшить результаты классификации. В данной работе ставилась задача сравнения качества работы нескольких алгоритмов и выявления возможностей дальнейшего совершенствования подхода на основе АФП, а не поиск наилучшего решения конкретной задачи классификации.

Поэтому было принято решение о применении простого метода бинаризации признаков:

1) Линейное нормирование признаков, то есть отображение в интервал $[0, 1]$;
2) Интервал $[0, 1]$ делился на 10 частей и каждому вещественному признаку соответствовал один интервал, в который он попал из 10 возможных.

Для вычисления минимальных гипотез был использован модифицированный алгоритм «Замыкай-по-одному» [2]. Согласно алгоритму процесс порождения формальных понятий начинается от понятий с наименьшими содержаниями к понятиям с наибольшими содержаниями (то есть «сверху-вниз» согласно порядку $\leqslant$ в решетке понятий).

Алгоритм порождает формальные понятия отдельно для каждого из классов до тех пор, пока не будут найдены все минимальные гипотезы.

Будем считать, что все признаки из M упорядочены и могут быть представлены своими номерами. Обозначим за $min(X)$ функцию, которая выдает объект из множества X с наименьшим номером, за $max(X)$ функцию, которая выдает объект с наибольшим номером. Переменная $prev(B)$ задает единственного предшественника формального понятия с содержанием B в решётке. Переменная $next\_i(B)$ обозначает номер следующего признака, который должен быть проверен как признак, порождающий потомка множества B. min_hyp обозначает список минимальных гипотез.

Алгоритм выделения минимальных гипотез:

$B := \varnothing,\ next\_i(B) := 1,\ prev(B) := \varnothing,\ min\_hyp := \varnothing$
пока $B \neq \varnothing$ или $next\_i(B) \leqslant |M|$
  пока $next\_i(B) \leqslant |M|$ и B'' не является гипотезой
    $i := next\_i(B)$
    если $min((B \cup \{i\})'' \setminus B) \leqslant i$ **то**
      $prev((B \cup \{i\})'') := B$
      $next\_i(B) := next\_i(B) + 1$
      $next\_i((B \cup \{i\})'') := min(\{j \mid i < j \,\&\, j \notin (B \cup \{i\})''\})$
      $B := (B \cup \{i\})''$
    иначе
      $next\_i(B) := next\_i(B) + 1$
  если B'' является гипотезой **то**
    $min\_hyp \leftarrow B''$
  $B := prev(B)$

Согласно алгоритму, процедура генерации «потомков» формального понятия прекращается, как только формальное содержание понятия является гипотезой. Таким образом вычисляются минимальные гипотезы, так как порождения других гипотез, подмножеством которых являются выделенные ранее гипотезы, не происходит.

Временная сложность вычисления всех минимальных положительных и отрицательных гипотез в худшем случае равна $O((|L(K_+) + |L(K_-)|) \cdot |M|^2 \cdot |G_+| \cdot |G_-|)$, $|L(K_+)|$ ($|L(K_-)|$) — размер решетки понятий контекста K_+ (K_- соответственно).

Однако стоит отметить, что результат алгоритма зависит от порядка предоставления признаков и выделенное множество гипотез может также содержать гипотезы, не являющиеся минимальными. Поэтому необходимо производить дополнительную фильтрацию списка min_hyp для удаления гипотез $h \in min\_hyp$, таких, что $\exists\, h_1 \in min\_hyp : h_1 \subset h$.

Эксперименты на реальных данных

Для тестирования алгоритма были использованы данные из UCI Machine Learning repository: [1]

1) *Liver Disorders (Заболевания печени).* Цель задачи — определить пациентов с заболеванием печени. Выборка содержит данные медицинских анализов крови 345 пациентов. Признаки представляют собой результаты анализов — 6 показателей: средний объем эритроцитов, щелочная фотосфаза, аланиновая трансаминаза (АЛТ), аспартатаминотрансфераза (АСТ), гаммаглютамилтранспептидаза, средний объём употребляемого алкоголя в день.
2) *Tic-tac-toe (Крестики-нолики).* Цель задачи — определить выигрышные ситуации. Выборка состоит из 958 игровых ситуаций, представляющих всевозможные конфигурации на поле в конце игры, где крестики ходят первыми. Признаки представляют собой каждый из квадратов на игровом поле (всего 9) и имеют значения «крестик» , «нолик» или «пусто».
3) *House-votes-84 (Голосование конгресса США в 1984 году).* Цель задачи — определить партию головавшего когрессмена (республиканец или демократ). Выборка состоит из результатов голосования 435 членов палаты представителей США. Признаками являются результаты голосования по 16 вопросам. Значениями является «высказался за», «высказался против», «воздержался».

Данные из задачи *Liver Disorders (Заболевания печени)* подверглись бинаризации согласно алгоритму, описанному в предыдущей главе. В итоге было сформировано 60 признаков.

Алгоритм распознавания реализован в среде MATLAB. Тестирование производилось с помощью скользящего контроля по 10 блокам. Результаты работы представлены в таблице 1. Также в таблице 1 указано время работы программы на системе с процессором Intel Core i5, 2.3 ГГц, 4 Гб ОЗУ.

Как видно из результатов тестирования, алгоритм в большинстве случаев отказывается от классификации по противоречию.

Модифицированный алгоритм

Для увеличения количества распознанных объектов была проведена модификация данного алгоритма. Первая модификация представляет собой отбор гипотез H, которым удовлетворяет некоторое количество объектов, большее чем параметр алгоритма $P : |Extent(H)| > P$. Это позволяет не учитывать шум и значительно уменьшает время работы алгоритма.

Далее были использованы следующие алгоритмы классификации на основе отобранных гипотез:

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

Таблица 1: Результаты тестирования простого алгоритма

Задача	*% отказа по недостатку*	*% отказа по противоречию*	*% верно*	*% ошибок*	*Время работы (сек.)*	*Кол-во гипотез*
1. Liver Disorders	13.53%	22.8%	42.8%	20.88%	12, 8	258
2. Tic-tac-toe	0%	99.3%	0.7%	0%	51	1887
3. House votes 84	0%	37.96%	62%	0.04%	390	1777

Таблица 2: Результаты тестирования модифицированного алгоритма

Задача	*Алгоритм классификации*	*% отказ*	*% верно*	*% ошибок*	*Время работы (сек.)*	*Кол-во гипотез*
1. Liver Disorders	АФП, голосование, $P = 0,3\%$	$0,28\%$	$65,5\%$	$34,1\%$	$9,71$	110
	SVM	0%	$60,7\%$	$39,30\%$	$0,11$	-
	$C4.5$	0%	$60,6\%$	$39,40$ %	$0,44$	-
2. Tic-tac-toe	АФП, метрика (а), $P = 3\%$	0%	100%	0%	$3,16$	17
	SVM	0%	$98,3\%$	$1,7\%$	$0,4$	-
	$C4.5$	0%	$89,1\%$	$10,90\%$	$0,16$	-
3. House votes	АФП, голосование, $P = 2,8\%$	$0,6\%$	$96,3\%$	$3,1\%$	$94,7$	510
	SVM	0%	$95,7\%$	$4,3\%$	$0,09$	-
	$C4.5$	0%	$93,1\%$	$6,9\%$	$0,17$	-

1) Простое голосование (голосование по большинству);
2) Введение метрики $\rho(\boldsymbol{x}, \boldsymbol{y})$:
   (а) $\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{|\boldsymbol{x} \cap \boldsymbol{y}|}{|\boldsymbol{y}|}$;
   (б) $\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{|\boldsymbol{x} \cap \boldsymbol{y}|}{|\boldsymbol{x} \cup \boldsymbol{y}|}$;
   (в) $\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{2} \cdot |\boldsymbol{x} \cap \boldsymbol{y}|}{|\boldsymbol{x}| + |\boldsymbol{y}|}$;
   (г) $\rho(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{|\boldsymbol{x} \oslash \boldsymbol{y}|}{|\boldsymbol{x} \cup \boldsymbol{y}|}$, $\boldsymbol{x} \oslash \boldsymbol{y} = (\boldsymbol{x} \backslash \boldsymbol{y}) \cup (\boldsymbol{y} \backslash \boldsymbol{x})$.

Результаты тестирования модифицированного алгоритма сравнивались с результатами алгоритмов классификации SVM и $C4.5$. В таблице 2 представлены лучшие результаты на скользящем контроле по 10 блокам.

Как видно из результатов тестирования, качество работы алгоритма на основе Анализа Формальных Понятий не уступает известным алгоритмам SVM и $C4.5$, а в лучшем случае

превосходит их. Однако алгоритмы на основе Анализа Формальных Понятий проигрывают по времени работы. Таким образом применение таких алгоритмов целесообразно в задачах, требующих точного решения и не имеющих сильного ограничения по времени решения.

Также стоит отметить, что отбор гипотез позволяет существенно улучшить качество распознавания. Процент верно распознанных объектов на стадии тестирования у алгоритма на основе голосования по большинству в среднем больше, чем у других алгоритмов на основе метрик (а)-(г).

Модифицированный алгоритм классификации эффективно решает задачи с номинальными и бинарными признаками. Решение задачи «Liver Disorders» может быть улучшено с помощью другой процедуры бинаризации признаков, например применения интервалов не фиксированной длины.

Заключение

В статье рассматриваются модификации простого алгоритма на основе АФП и сравнивается их эффективность.

Установлено, что при введении ограничения на размер формального объема гипотез и применения метода голосования по большинству классификатор способен эффективно решать задачи распознавания с номинальными и бинарными признаками. Также благодаря порождению всевозможных минимальных гипотез значительно уменьшается количество отказов от классификации по сравнению с алгоритмами, описанными в [1].

Анализ Формальных Понятий представляет собой перспективное направление для исследований в области обработки данных и извлечения зависимостей.

Дальнейшие исследования могут быть направлены на решение задач классификации с вещественными признаками и применения методов АФП для предобработки признакового пространства, в частности для повышения компактности данных. Также важной задачей является реализация быстрых алгоритмов построения решетки формальных понятий.

Литература

[1] *Онищенко А. А., Гуров С. И.* Классификация на основе АФП и бикластеризации: возможности подхода // Прикладная математика и информатика: Труды факультета Вычислительной математики и кибернетики. — 2011. — Т. 38. — С. 77–87.

[2] *Kuznetsov S.* Mathematical aspects of concept analysis. // Journal of Mathematical Science. — 1996. — Vol. 80, No. 2. — Pp. 1654–1698.

[3] *Meddouri N., Meddouri M.* Classification Methods based on Formal Concept Analysis // CLA 2008 (Posters), Palackŕy University, Olomouc, 2008. —Pp. 9–16.