Обучение на помеченных графах и их проекциях

С.О. Кузнецов[1], М.В.Самохин[2]

В работе исследуется методы представления и анализа данных, задаваемых помеченными графами. В качестве экспериментального материала выбрано несколько задач о прогнозировании биологической активности химических соединений, а в качестве методов прогноза – 4 метода машинного обучения. Помеченные графы представляются без потерь с помощью бинарных признаков на основе техники порядкового шкалирования, предлагается метод приближенного представления графов проекциями, а также техника редуцирования объектно-признаковой матрицы. Эти приемы позволяют уменьшать размерность данных в несколько раз без существенного влияния на качество прогнозов.

1. Постановка задачи

В последние годы проблематика обучения на данных, заданных помеченными графами привлекает все большее внимание среди исследователей в области машинного обучения и разработки данных [14, 15, 16, 8, 13, 22, 23, 24, 11]. В данной работе мы подходим к решению данного вопроса, используя методы порождения замкнутых множеств помеченных графов и их приближений. С одной стороны этот подход основан на вычислении наиболее частных (наименее общих) обобщений положительных и отрицательных примеров, который зарекомендовал себя в реальных практических приложениях, в том числе в открытом соревновании по предсказательной токсикологии [7, 20]. С другой стороны, порождение (частых) замкнутых подмножеств признаков оказалось полезным для вычисления множества ассоциативных правил, имеющих хорошую поддержку [18]. Этот факт объясняет недавний рост интереса к вычислению замкнутых графов в области разработки данных [25]. В работе [25] алгоритм Close-Graph, порождающий замкнутые графы, вычисляет все "частые" графы намного быстрее чем предыдущая программа gSpan [24] тех же авторов, а также ILP-программа WARMR [14].

Основным языком описания данных в ИС типа ДСМ для задач, связанных с химией, является Фрагментарный Код Суперпозиций Подструктур (ФКСП) [1], в котором каждое химическое соединение

[1] *125190, Москва, ул. Усиевича 20, ВИНИТИ РАН, serge@viniti.ru*
[2] *125190, Москва, ул. Усиевича 20, ВИНИТИ РАН, samohin_m@mtu-net.ru*

представляется множеством дескрипторов. С точки зрения вычислений достоинства языка состоят в простоте реализации теоретико-множественных операций и отношений над объектами. Однако, существуют задачи, в которых код ФКСП является недостаточным для описания структуры объекта, как, например, в задаче прогнозирования путей метаболизма химических соединений в организме. Поэтому целесообразно использовать более точное описание исходных данных (например, описание графами) в тех случаях, когда коды ФКСП не дают необходимых результатов.

В этой работе обсуждается подход, применяемый для анализа данных, представленных помеченными графами и их проекциями. Для описания графов в объектно-признаковом виде используется порядковое шкалирование - стандартная техника теории Анализа Формальных Понятий (АФП) [10]. Помеченные графы представляются без потерь с помощью бинарных признаков на основе техники порядкового шкалирования, однако первоначально число порождаемых признаков описания может быть очень велико. В связи с этим предлагается метод приближенного представления графов проекциями, а также техника редуцирования объектно-признаковой матрицы, которые позволяют уменьшать размерность данных в несколько раз без существенного влияния на качество результатов. Сравниваются результаты прогноза биоактивности химических соединений, получаемые разными методами машинного обучения: ДСМ-методом, методом порождения деревьев решений C4.5, наивным методом Байеса (в котором считается, что структурные признаки распределены независимо), а также методом JRip, порождающего правила с исключениями. Массивы, на которых проводились сравнения методов – это стандартный массив PTC [20], массив галогенозамещенных углеводородов [6], массив спиртов [2] и массив полициклических ароматических углеводородов (ПАУ) [4].

2. Основные определения

В работе [3] была рассмотрена модель обучения из [5] в терминах АФП. В этой модели предполагается, что причина наличия целевого свойства связана с теми общими признаками, которые содержат описания всех объектов, обладающих этим свойством. В работе [9] были введены обобщенные определения ДСМ-гипотез для данных, на которых задана полурешеточная операция сходства.

Вкратце обучение гипотезам «с запретом на контрпример» выглядит следующим образом: для целевого признака имеются множества G_+ и G_- положительных и отрицательных примеров с описаниями, на которых задана полурешеточная (т.е. идемпотентная, коммутативная и ассоциативная) операция сходства $*$ и соответствующее ей отношение

вложения $\sqsubseteq$. Для положительных примеров $g_1,...,g_n$ и их описаний $\delta(g_1),\ldots, \delta(g_n)$, соответственно, пересечение описаний $h = \delta(g_1) * \ldots * \delta(g_n)$ положительных примеров $g_1,...,g_n$ есть положительная гипотеза (с запретом на контрпример), подтверждаемая примерами $g_1,...,g_n$, если h не вкладывается в описание ни одного отрицательного примера g_-, т.е. h o $\delta(g_-)$. Аналогично, определяются отрицательные гипотезы. Гипотезы используются для классификации примера g_τ с неопределенным значением целевого свойства следующим образом: если существует такая положительная гипотеза h_+, что h_+ m g_τ, то g_τ классифицируется положительно. Отрицательная классификация определяется аналогично, возможны также противоречивая классификация и неопределенная классификация (детали см. в [9]).

Рассмотрим следующую полурешетку $(G, *)$ на множествах графов с метками вершин и ребер из некоторого множества L. Определение полурешетки начнем с определения частичного порядка на графах. В простейшем случае порядок задается отношением изоморфизма подграфу: $\Gamma_2 \le \Gamma_1$, если Γ_2 изоморфен подграфу Γ_1. В общем случае, когда задан порядок на метках, порядок на графах должен учитывать еще и порядок на метках.

Операция пересечения * на множествах графов может быть задана следующим образом: для пары графов X и Y из G, $\{X\} * \{Y\} := \{Z \mid Z \le X, Y, \ \forall Z_* \le X, Y \ Z_* \times Z\}$ (множество всех максимальных общих подграфов графов X и Y). В общем случае, $\{X_1,...,X_k\} * \{Y_1,...,Y_m\} := \mathrm{MAX}_{\le}(+_{i,j}(\{X_i\} * \{Y_j\}))$, где MAX($Z$) выбирает максимальные элементы из множества Z (детали см. в [9]).

Данное определение сходно с определением замкнутого графа [25], используемого для вычисления ассоциативных правил на парах графов. Замкнутые графы определяются в работе [25] (в терминах так называемого “counting inference”) следующим образом:

Для массива данных D (содержащего данные в форме помеченных графов), поддержка графа g или support(g) есть множество (или размер множества) графов в D, у которых есть подграф, изоморфный графу g.

Граф g является замкнутым, если все надграфы f графа g имеют другую поддержку. Подчеркнем, что в данном определении речь идет о замкнутом графе, а не замкнутом множестве графов. Как уже говорилось выше, замкнутые множества графов образуют нижнюю полурешетку по отношению к оператору *, а всякая конечная нижняя полурешетка может быть пополнена до решетки введением единичного (максимального) элемента. Нам не известно подобной операции на замкнутых графах (по-видимому, такая операция не может быть определена) в частично-упорядоченном множестве замкнутых графов, в общем случае,

существуют множественные инфинумы и супремумы двух замкнутых графов. Тем не менее, существует тесная связь между понятием замкнутого графа и понятием замкнутого множества графов.

Утверждение.

Пусть массив данных, задан узорной структурой $(E, (D, \sqcap), \delta)$. Тогда

1. Для любого замкнутого графа g существует замкнутое множество графов G такое, что $g \in G$.

2. Для замкнутого множества графов G любой граф $g \in G$ есть замкнутый граф.

В силу ограниченности размера публикации мы не приводим здесь определений стандартных методов обучения C4.5, Naïve Bayes, JRip и отсылаем читателя к книге по коллекции методов машинного обучения WEKA на языке Java [23].

Так как проверка отношения $\leq$ на помеченных графах является NP-полной задачей, в [9] были определены операторы проекций, используемые для приближенного описания. Отображение $\psi : G \to G$ называется оператором проекции, если выполнены следующие условия: 1) если $x \sqsubseteq y$, то $\psi(x) \sqsubseteq \psi(y)$ (монотонность), 2) $\psi(x) \sqsubseteq x$ (сжатие), и 3) $\psi(\psi(x)) = \psi(x)$ (идемпотентность).

Любая проекция φ полной полурешетки $(G, \sqcap)$ сохраняет операцию $\sqcap$, т.е., для любых $X, Y \in G$, $\psi(X \sqcap Y) = \psi(X) \sqcap \psi(Y)$, что позволяет установить связь между гипотезами и классификациями в исходном представлении и в его приближенном описании, полученном с применением оператора проекции.

Теперь мы рассмотрим применение модели обучения из Раздела 2 для решения задачи предсказания биологической активности химических соединений. В наших экспериментах мы использовали данные, представленные на соревновании Predictive Toxicology Challenge (PTC) [20]. Исходным представлением данных для массива PTC были молекулярные графы - помеченные графы. Такое представление позволяет применять методы, описанные в Разделе 2.

Можно сказать, что изначально соединения описаны одним признаком «граф», который сопоставляет каждому химическому соединению его молекулярный граф. Применяя шкалирование к исходному описанию, мы получаем новое описание соединений множеством признаков, в качестве которых выступают все связные подграфы графов, рассматриваемых в исходном описание, а каждому графу приписываются все содержащиеся в нем подграфы в качестве признаков.

Сложность процедуры порождения всех подграфов произвольного графа связана со сложностью проверки изоморфизма графу и

изоморфизма подграфу. Существует несколько известных алгоритмов для решения этих задач, в частности, алгоритм B.D. McKay [17] для проверки изоморфизма графу и алгоритм J.R. Ullmann [21] для проверки изоморфизма подграфу. В связи со сложностью указанных проверок в нашем подходе мы используем *k*-проекции исходных графов. Идея проекции, описанная в Разделе 2.1 для общей полурешетки, может быть уточнена для случая полурешетки на графах следующим образом:

Определение 1. *Пусть* $\Gamma = ((V,l), (E,b))$ - *помеченный граф. Множество* $S_\Gamma = \{\Gamma_* = ((V_*, l_*), (E_*, b_*)) \mid \Gamma_*$ *- связный граф*, $\Gamma_* \le \Gamma$, $|V_*| \le k\}$ *будем называть k-проекцией графа* Γ.

Таким образом, *k*-проекция помеченного графа Γ есть множество всех подграфов (с точностью до изоморфизма) графа Γ с числом вершин не более чем *k*. Аналогично, *k*-проекция множества помеченных графов определяется через объединение проекций графов из этого множества (из этого объединения необходимо удалить немаксимальные по отношению ≤ графы).

После преобразования графов в объектно-признаковое представление и вычисления проекций, получается несколько массивов данных, соответствующих различным приближенным представлениям (*k*-проекциям) исходного массива данных.

3. Экспериментальные результаты

Целью наших экспериментов была проверка возможностей нашего метода в сравнение с результатами, полученными другими участниками в [19]. Для проведения экспериментов программный модуль, реализующий идеи, описанные в Разделе 3, был интегрирован в систему интеллектуального анализа данных QuDA [12].

3.1. Использование системы QuDA для предсказания биологической активности химических соединений: метод последовательного покрытия

В программной системе QuDA [12] реализованы несколько вариантов модели обучения, описанной в Разделе 2. В наших экспериментах на массиве PTC [20] мы использовали стратегию последовательного покрытия на основе структурного сходства. Эта процедура не порождает множества всех гипотез (в отличие от стратегии, обычно используемой в ДСМ-системах, порождающей все минимальные гипотезы), а последовательно порождает некоторые минимальные гипотезы, позволяющие объяснить все примеры из обучающего множества. В некотором смысле такая стратегия лежит посередине между стратегией с

ДСМ-гипотезами (с запретом на контрпример) и методом порождения деревьев решений. Хотя эта стратегия имеет некоторый недостаток: зависимость от выбранного порядка на объектах, она отличается вычислительной эффективностью и хорошей точностью приближения к «полной» стратегии (с ДСМ-гипотезами с запретом на контрпример) на многих реальных массивах данных.

Одним из общепринятых методов, используемых в машинном обучении, для сравнения работы алгоритмов классификации является ROC-анализ [19]. В данной работе мы использовали ROC-диаграммы для изучения проекций графов.

3.2. Результаты экспериментов на различных массивах данных

Кратко перечислим результаты, которые были получены с использованием проекций на различных массивах данных.

Подробное описание массива данных Predictive Toxicology Challenge [19] и результатов, полученных с использованием ДСМ-системы и описанием исходных данных кодами ФКСП даны в работе [7].

Особенности нашего эксперимента с проекциями графов: 1) язык описания - *k*-проекции (значение *k* варьировалось от 1 до 8); 2) использование стратегии порождения некоторой части всего множества гипотез (стратегия последовательного покрытия, описанная в Разделе 4.1).

Для группы «самцы крыс» использование 4-, 5-, 6-, и 8-проекций, а для группы «самки крыс» использование 5- и 7-проекций стало одними из четырех «новых» лучших методов для данной группы. Использование 4-проекций ведет к достаточно хорошей классификации, соответствующая точка лежит выше «старой» ROC-кривой.

В [2] были описаны результаты исследований соотношения между структурой различных спиртов и острой токсичностью для крыс и мышей с использованием ДСМ-метода и ФКСП-кодов. Используя в качестве языка описания *k*-проекции, были проведены эксперименты с применением различных методов машинного обучения. Полученные результаты показали, что использование *k*-проекций вместо ФКСП-кодов в качестве языка представления позволяет увеличить число правильных предсказаний и уменьшить число ошибочных предсказаний для всех четырех методов машинного обучения (ДСМ, C4.5, Naïve Bayes, Jrip).

В следующем эксперименте изучалась канцерогенность галогензамещенных алифатических углеводородов [6]. Эксперимент состоял из двух частей. В первой части классифицировались соединения относительно целевого признака «быть канцерогенным». Результаты экспериментов показали, что использование *k*-проекций (как и в случае с предыдущими экспериментами) позволяет улучшить результаты, полученные с использованием ФКСП-кодов. Во второй части

исследовалась канцерогенность непрямых канцерогенов (соединений, канцерогенная активность которых определяется их биоактивацией в организме) на массиве данных из [6]. Для этой задачи схема, основанная на чисто структурном описании молекул, не всегда давала правильные результаты. Поэтому ко всем соединением из выборки были добавлены числовые значения (параметры, характеризующие энергию активации реакции метаболизма в организме), на которых операция сходства была задана полурешеткой замкнутых интервалов. В этом случае полурешетка сходства объектов задается парами вида *(множество графов, числовой интервал)*. Использование числовых параметров с ФКСП-кодами и *k*-проекциями вместе с различными методами машинного обучения показало (как в предыдущих экспериментах), что описания с помощью *k*-проекций приводят к большому числу правильных предсказаний и меньшему числу ошибок, чем описания с помощью ФКСП-кодов. В тоже время ДСМ-метод дает наименьшее число ошибок в классификации в сравнении с предсказаниями, полученными с помощью других методов машинного обучения на тех же описаниях, а по соотношению полнота/точность предсказания, наилучшим для данного массива представляется метод C 4.5.

Последним экспериментом, на котором проводилась проверка возможностей описанного выше метода, был эксперимент по исследованию канцерогенности полициклических ароматических углеводородов (ПАУ). Соединения, которые исследовались в [4], были описаны с помощью *k*-циклических проекций (для множества молекулярных графов было построено множество подграфов, состоящих из *k* связанных циклов из минимального циклического базиса хотя бы одного из исходных графов). Согласно биохимической и квантово-химической модели [4] канцерогенная активность ПАУ определяется их биоактивацией с образованием метаболитов, способных связываться с ДНК. Поэтому, как в эксперименте с галогензамещенными алифатическими углеводородами, вместе со структурной частью рассматривался числовой параметр, характеризующий энергию образования наиболее активного конечного метаболита ПАУ. Проведенные эксперименты показали, что применение ДСМ-метода на данном массиве дает лучшие результаты среди всех рассматриваемых методов машинного обучения. В тоже время использование проекций ведет к незначительному увеличению числа правильных предсказаний по отношению к результатам, полученным с применением ФКСП-кодов.

4. Заключение и дальнейшая работа

В работе проводилось сравнение классификации данных ряда массивов с использованием представления соединений в виде графов и в

виде ФКСП с помощью четырех методов машинного обучения. Результаты компьютерных экспериментов позволяют утверждать о состоятельности подхода, основанного на операторе проекции графов, поскольку он значительно сокращает временные затраты на анализ данных обучающей выборки и последующее применение полученных гипотез, с целью классификации, к данным тестовой выборки, в тоже время не приводя к большим информационным потерям (т.е. неправильным классификациям). Вместе с тем, эксперименты показывают оправданность применения техники редуцирования контекстов. Редуцирование контекста не приводит к изменениям результатов ДСМ-метода, приводит к несущественным изменениям результата метода порождения деревьев решений (5-10%) и приводит к малым изменениям (5-20%) для наивного метода Байеса и JRip.

Представляется осмысленным применение техники проекции для других задач, использующих графовое представление данных, таких как анализ графов конфликтов, классификация документов в поисковых машинах Интернет и др.

Список литературы

[1] Блинова В.Г. и Добрынин Д.А. Языки представления химических структур в интеллектуальных системах для конструирования лекарств // НТИ сер.2, №6, 2000.

[2] Блинова В.Г., Добрынин Д.А., Жолдакова З.И. и Харчевникова Н.В. Изучение соотношений структура-токсичность спиртов с использованием ДСМ-метода // НТИ, сер. 2, № 10, 2001.

[3] Кузнецов С.О. и Финн В.К. О модели обучения и классификации, основанной на операции сходства // Обозрение Прикладной и Промышленной Математики 3, №1, 1996.

[4] Максин М. В., Харчевникова Н.В., Блинова В.Г., Добрынин Д.А. и Жолдакова З.И. Система, реализующая комбинаторно-численный подход к проблеме прогноза свойств химических соединений. Прогноз канцерогенности полициклчиеских ароматических углеводородов (ПАУ) // НТИ, сер. 2, №1, 2004

[5] Финн В.К., Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники. Сер. Информатика. - М.: ВИНИТИ, 1991.

[6] Харчевникова Н.В., Блинова В. Г., Добрынин Д.А., Максин М.В. и Жолдакова З.И. Применение ДСМ-метода и квантово-химических расчетов для прогноза канцерогенности и хронической токсичности галогензамещенных алифатических углеводородов // НТИ, сер. 2, №12, 2003.

[7] Blinova V.G., Dobrynin D.A., Finn V.K., Kuznetsov S.O., and Pankratova E.S., Toxicology analysis by means of the JSM-method // Bioinformatics, №19, 2003.

[8] Borgelt C. and Berthold M.R. Mining Molecular Fragments: Finding Relevant Substructures of Molecules. Proc. *2nd IEEE International Conference on Data Mining*, *ICDM'02*. N. Zhong and P.S. Yu, Eds., IEEE Press, 2002.

[9] Ganter B. and Kuznetsov S.O., Pattern Structures and Their Projections, Proc. *9th Int. Conf. on Conceptual Structures, ICCS'01*, G. Stumme and H. Delugach, Eds., Lecture Notes in Artificial Intelligence, 2120, 2001.

[10] Ganter B. and Wille R. *Formal Concept Analysis. Mathematical Foundations*, Springer, 1999.

[11] Gonzalez J.A., Holder L.B. and Cook D.J. Application of Graph-Based Concept Learning to the Predictive Toxicology Domain. Proc. *Workshop on Predictive Toxicology Challegnge at the 5th Conference on Data Mining and Knowledge Discovery*, PKDD'01. C. Helma, R.D. King, S. Kramer, and A. Srinvasan, Eds. http://www.predictive-toxicology.org/ptc/, 2001, September 6.

[12] Grigoriev P.A., Yevtushenko S.A. and Grieser G. QuDA, a data miner's discovery environment // Technical Report AIDA 03 06, FG Intellektik, FB Informatik, Technische Universität Darmstadt, http://www.intellektik.informatik.tu-darmstadt.de/~peter/QuDA.pdf, September 2003.

[13] Inokuchi A., Washio T. and H. Motoda. Complete Mining of Frequent Patterns from Graphs: Mining Graph Data // Machine Learning, 50(3), 2003.

[14] King R.D., Srinivasan A. and Dehaspe L. WARMR: A Data Mining tool for chemical data. // J. of Computer-Aided Molecular Design, 15(2), 2001.

[15] Kramer S. Structural Regression Trees. Proc. *13th National Conference on Artificial Intelligence*, *AAAI-96*, Cambridge/Menlo Park, 1996.

[16] Kuznetsov S.O. Learning of Simple Conceptual Graphs from Positive and Negative Examples. Proc. *Principles of Data Mining and Knowledge Discovery, Third European Conference*, *PKDD'99*. J. Zytkow and J. Rauch, Eds., Lecture Notes in Artificial Intelligence, 1704, 1999.

[17] McKay B.D. Practical graph isomorphism // Congressus Numerantium, №30, 1981.

[18 Pasquier N., Bastide Y., Taouil R. and Lakhal L. Efficient Mining of Association Rules Using Closed Itemset Lattices // J. Inf. Systems, 24(1), 1999.

[19] Provost F. and Fawcett T. Robust classification for imprecise environments // Machine Learning, №42, 2001.

[20] http://www.predictive-toxicology.org/ptc/.

[21] Ulmann J.R. An algorithm for subgraph isomorphism // J. of Assoc. Comput. Mach., №23, 1976.

[22] Washio T. and Motoda H. State of the art of graph-based data mining // SIGKDD Explorations Newsletter, 5(1), 2003.

[23] Witten I.H. and Frank E. Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, 2000

[24] Yan X. and Han J. gSpan: Graph-Based Substructure Pattern Mining. Proc. *IEEE Int. Conf. on Data Mining*, *ICDM'02*. IEEE Computer Society, 2002.

[25] Yan X. and Han J. CloseGraph: mining closed frequent graph patterns. *Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, *KDD'03.* L. Getoor, T.E. Senator, P. Domingos and C. Faloutsos, Eds. ACM Press, 2003.