Эмпирическая теория интеллектуального анализа данных

Загоруйко Н.Г.

Институт математики СО РАН

1. Введение

Главным отличительным свойством систем, обладающих интеллектом, является умение делать правильные предсказания. Такие предсказания становятся возможными, только если в потоке событий удается обнаруживать закономерные связи между наблюдаемыми ситуациями. В результате возникает система приблизительно верных знаний об окружающем мире, которые фиксируются в качестве эмпирических гипотез или теорий [1–3]. Слово «эмпирических» говорит о том, что такие теории и гипотезы могут подтверждаться или опровергаться только результатами наблюдений за событиями реального мира.

Хорошие теории удовлетворяют критерию Д.И. Менделеева, говорившему: «Цель науки – истина и польза». Истинность эмпирической теории состоит в высокой степени **подтвержденности** описываемых закономерностей результатами многочисленных наблюдений. А польза теории заключается в наличии обоснованных, без нее не очевидных **рекомендаций** о том, что и как нужно делать в конкретной ситуации.

2. Что такое эмпирическая теория?

В формальном представлении эмпирическая теория, или, что то же самое, эмпирическая гипотеза h, может выглядеть так: $h=<W,O,V,T>$.

Здесь W – множество объектов или явлений, о котором теория делает свои утверждения. Например, «все материальные тела» или «все методы распознавания образов».

Символ O – инструкция о том, чем и как проводить наблюдения, чтобы они относились к рассматриваемой теории, и на каком языке V записывать результаты наблюдений. Инструкция должна давать ответ на вопрос: данное наблюдение получено в соответствии с этой инструкцией или нет? Кроме того, любое наблюдение, проведенное в соответствии с ней, должно допускать описание результатов в виде протокола prv в словаре языка V.

Символом T обозначен «тестовый» алгоритм, который работает с любым протоколом на языке V и принимает одно из двух значений: $T(pr^v) = 1$, если протокол согласуется с гипотезой, или $T(pr^v) = 0$ в противном случае. На изоморфных протоколах (протоколах, имеющих одинаковое эмпирическое содержание) алгоритм T принимает одинаковые значения.

Эмпирический смысл теории h определяется следующим соглашением: *Мир W таков, что если мы будем наблюдать его способом O, то никогда не получим протокола prv, на котором тестовый алгоритм примет значение* $T(pr^v)=0$. Эмпирические теории описывают потенциально опровержимые гипотезы об устройстве изучаемого мира. Если h не опровергается протоколом pr^v, то считается, что теория h согласуется с данным наблюдением, а протокол pr^v подтверждает теорию. Теория h никогда не может быть доказана раз и навсегда, но может быть опровергнута одним единственным достоверным экспериментом.

Эмпирические теории, касающиеся объектов и явлений некоторой прикладной области, могут описывать закономерные связи между разными подмножествами их характеристик. Так что развитая теория H сложной эмпирической системы может включать в свой состав не одну, а несколько эмпирических теорий типа h. Например, в области DM можно сформулировать свою теорию h для методов решения задач каждого типа: теорию h_s методов автоматической классификации, теорию h_x методов выбора информативных признаков и т.д. Наряду с этим, можно построить теорию H, описывающую некоторые общие свойства методов решения задач DM разного типа. В дальнейшем термин «эмпирическая гипотеза» будем применять к теориям, касающимся методов решения задач одного типа, а термином «эмпирическая теория» будем называть теорию DM, описывающую общие свойства методов решения любых задач Data Mining.

Если отвлечься от эмпирического содержания теорий и гипотез, то можно отметить такие их свойства, как потенциальная опровержимость Q, степень подтвержденности P, глубина объясненности E, простота S и красота B формулировки [2]. **Закономерностью** R называется эмпирическая гипотеза h, сопровождаемая значениями ее характеристик: $R = <h, Q, P, E, S, B>$. Развитие теории связано с усилением гипотез, с улучшением этих характеристик. Наиболее важной является характеристика потенциальной опровержимости Q. Улучшение Q связано с переходом, например, от неопровержимого, а потому и бессодержательного, утверждения типа «В этом мире все возможно», до очень рискованного, но пока не опровергнутого утверждения «Сила равна массе, умноженной на ускорение». Чем больше воображаемых протоколов могли бы опровергнуть гипотезу, тем выше величина Q и тем полезнее она для практики. Наше доверие к гипотезе растет также с увеличением количества P подтвердивших ее экспериментов. Большое значение имеют ответы на вопросы «Как?» и «Почему?», углубляющие объясненность E гипотезы [4]. В истории науки имеется много

примеров, показывающих полезность простых (S) и красивых (B) формулировок гипотез.

Из сказанного выше видно, что для построения эмпирической теории некоторой прикладной области требуется определить, какие объекты W этой области и какие их свойства мы будем изучать средствами O и записывать на языке V. Кроме того, нужно описать алгоритм T, с помощью которого будем определять, что в этой области возможно и чего не может быть. Попробуем применить этот подход к предметной области, которая связана с Data Mining, т. е. с методами автоматического обнаружения эмпирических закономерностей и их использованием для выработки предсказаний.

3. Состояние проблемы DM

Проблема обнаружения эмпирических закономерностей является одной из центральных в Искусственном Интеллекте. Математические методы, используемые для обнаружения закономерностей (знаний), называются методами Интеллектуального Анализа Данных или Data Mining. Именно эти методы и являются объектами W нашего рассмотрения. Среди задач DM, указанных в их классификации [2], наибольшей популярностью пользуются задачи следующих типов:

1. Задача типа S: создание классификационной структуры заданного множества M объектов. Множество M делится алгоритмами классификации на k классов (кластеров, таксонов, образов) по похожести характеристик объектов. Получаемые классификации могут иметь одноуровневую или многоуровневую иерархическую структуру. Границы классов могут быть описаны простыми линейно разделимыми фигурами или фигурами произвольной сложности.

Эмпирическая гипотеза методов автоматической классификации hs может выглядеть так:

$$hs = <W, O, V, T>,$$

где W – все возможные методы $w_1, w_2, \ldots, w_j, \ldots w_J$ разбиения конечного множества M объектов на k непересекающихся непустых подмножеств (кластеров) $S_1, S_2, \ldots, S_k$. Значение k может находиться в заданных пределах от k_{min} до k_{max}.

O – способы оценки и записи на языке V характеристик получаемых вариантов разбиений. Среди характеристик кластеризации обычно отмечаются такие «геометрические» характеристики, как мера похожести объектов одного кластера друг на друга или на эталонный объект, удаленность кластеров друг от друга и т. д. Помимо этого следует обращать внимание и на «индуктивные» свойства получаемой классификации, чтобы по известным свойствам объектов кластера можно было определять и некоторые другие свойства этих объектов. Это требование предъявляется к алгоритмам т. н. «естественной» классификации [5, 6].

T – тестовый алгоритм, который получает очередной протокол pr^v, оценивает интегральную характеристику качества F описанной в протоколе кластеризации и возвращает решение 1, если качество превышает заданный порог F^*, и 0 – в иных случаях.

Такая гипотеза позволяет разделить все возможные методы кластеризации на допустимые (W^+) и недопустимые (W^-). Если множество всех возможных алгоритмов классификации W остается постоянным, то усиление гипотезы h_s может состоять в расширении набора характеристик O, влияющих на качество кластеризации, и в повышении порога качества F^* в тестовом алгоритме. Алгоритм T можно устроить так, чтобы он сравнивал между собой несколько протоколов pr^v и выдавал 1 только одному из них, полученному самым лучшим алгоритмом классификации. Потенциальная опровержимость Q такой гипотезы, утверждающей, что некий алгоритм w_j лучше других, пропорциональна количеству конкурирующих алгоритмов. Степень подтвержденности P тем выше, чем больше сравнительных испытаний выиграл алгоритм w_j. Характеристика E будет тем больше, чем глубже будет объяснено, как устроен алгоритма w_j и почему он работает лучше, чем другие алгоритмы. Будет полезно, если алгоритм будет описан просто и изящно.

Аналогичным способом может быть описано содержание всех других задач DM и соответствующих гипотез, касающихся методов их решения. Далее мы ограничимся только описанием содержания задач разных типов.

2. Задача типа X: формирование системы X информативного описания объектов множества M, предварительно разделенного на k классов. Система из N первичных признаков задается экспертами. При решении этой задачи методом фильтрации («filtering») из N признаков выбирается наиболее информативное подмножество из $n < N$ признаков. При методе селекции («selection») N первичных признаков преобразуются в более информативную систему из n вторичных признаков [7].

Для оценки информативности признаков используются критерии двух видов: косвенные (Indirect) и прямые (Wrapping). К числу косвенных относится оценка энтропии плотности распределений образов в отдельных частях пространства признаков, сложность структуры логических решающих правил, расстояние между образами, деленное на сумму их дисперсий (критерий Фишера) и т. д. Считается, что прямые методы (One-Leave-Out, Cross-Validation и т. д.) являются более трудоемкими, но и более надежными критериями информативности. Процент правильно распознанных контрольных объектов, не участвовавших в обучении, принимается в качестве ожидаемой надежности будущего распознавания реальной контрольной выборки.

3. Задача типа D: по обучающей выборке, в которой каждый из M объектов отнесен к одному из k классов, строятся решающие правила D, по которым эта выборка распознается с заданной

надежностью. Как и в предыдущем случае, здесь используются косвенные и прямые методы оценки качества построенных решающих функций.

4. Задача типа *Z*: заполнение пробелов («анализ некомплектных данных» или «Inserting») и обнаружение ошибок в таблицах данных («очистка данных» или «Cleaning») [2]. Обнаруживаемые закономерные связи между разными частями таблицы данных используются для предсказания наиболее правдоподобного значения пропущенного или искаженного элемента. Ожидаемая ошибка заполнения пробела оценивается прямым методом – по критерию минимума ошибок предсказания известных элементов таблицы. Задача этого типа ставится и при необходимости обнаружения отклонений от нормы (в частности, обнаружение мошенничества – «fraud detection»). При этом требуется сформировать описание объекта или процесса «в норме» и научить программу своевременно обнаруживать отклонения от нормы.

5. Задача типа *P*: прогнозирование динамических процессов. По данным, описывающим наблюдаемую динамику развития процесса, требуется обнаружить закономерную связь прошлого с будущим и затем предсказать характеристики процесса в заданные будущие моменты времени. Критерием качества, который применяется при обучении алгоритма, служит минимум ошибок прогнозирования при ретроспективном анализе.

6. Задача типа *A*: обнаружение ассоциаций. Нужно найти устойчивые связи между значениями одного подмножества характеристик X_1 и другого подмножества характеристик X_2. Затем по известным значениям X_1 требуется предсказывать значения характеристик X_2. И здесь об ожидаемой ошибке будущих решений обычно судят по результатам ретроспективного анализа.

Кроме этих задач основных типов встречаются задачи комбинированного типа [2]. Например, задача таксономии с одновременным выбором наиболее информативного подпространства признаков (**задача типа** *SX*). Или еще более сложная задача одновременного построения классификации, выбора признаков и построения решающего правила (**задача типа** *SDX*). Ситуация дополнительно усложняется, если данные описаны разнотипными признаками. Многие реальные задачи плохо статистически обусловлены: количество признаков бывает сравнимо и даже превышает количество объектов.

Такое большое разнообразие и сложность задач DM и их высокая актуальность привели к тому, что за последние 30–40 лет разработано большое количество алгоритмов для решения каждой из них. Попытка построить онтологию задач и методов DM [8] показала отсутствие единого подхода к решению не только всех этих задач, но и каждой задачи в отдельности.

4. Ориентиры дальнейшего развития

На что следует ориентироваться при поиске единого подхода к решению разных задач DM? Отметим, что все методы DM в той или иной мере имитируют способность человека систематизировать окружающий мир, формировать классификации, выбирать наиболее важные характеристики классов, определять принадлежность новых объектов к тому или иному классу. Отсюда ясно, что дальнейшее развитие методов DM следует искать на пути сближения свойств формальных моделей со свойствами человеческого механизма ориентации в окружающем мире. На какие особенности этого механизма следует обратить внимание?

1. Природный механизм основан на некотором универсальном подходе, позволяющем легко сочетать результаты решения разных задач в единую непротиворечивую цепочку решений. Так, разумные естественные классификации (задача таксономии *S*) основаны на использовании небольшого количества характеристик (задача выбора признаков *X*), по которым легко и надежно можно распознавать принадлежность объекта к своему классу (задача распознавания *D*).

Отсюда вытекает требование **согласованности** методов решения разных задач друг с другом. Следует оговориться, что требование согласованности результатов может быть обеспечено рядом существующих методов. Если комбинированные задачи решаются с помощью единого, например, алгебраического подхода [9] или логических деревьев [10], то результаты будут согласованы между собой. Но нельзя допускать, чтобы, например, классы, полученные методом *k*-means и разделяемые друг от друга произвольно ориентированными гиперплоскостями, распознавались с помощью логических решающих правил (деревьев), в которых используются плоскости, перпендикулярные осям координат.

2. Если законы распределения классов известны, то совместимость можно обеспечить, применяя методы, ориентированные на этот тип распределений. К сожалению, о характере распределений обычно ничего не известно. Несмотря на это, люди, в том числе и те, которые даже не подозревают о существовании законов распределений, постоянно решают такие задачи. То же относится и к настройке методов на такие особенности реальных задач, как характер зависимостей между признаками и соотношение между количеством объектов и признаков. Человеческий механизм заранее готов ко всем этим проявлениям реального мира. Отсюда вытекает требование **инвариантности** методов DM по отношению к виду законов распределений, характеру зависимостей между признаками и статистической обусловленности задачи.

3. Ясно, что при устремлении условий задачи к благоприятным для применения оптимальным методам (например, при нормальных распределениях, малом количестве признаков и большом

количестве обучающих объектов), совместимые и инвариантные методы должны давать решения, приближающиеся к оптимальным. Отсюда вытекает требование **потенциальной оптимальности** методов DM.

4. Хорошо известно, как сильно зависят результаты DM от таких факторов, как представительность выборки и независимость попадания объектов в обучающую выборку. Но, к сожалению, мы никогда не сможем узнать, все ли разнообразие объектов класса представлено в обучающей выборке и достаточно ли полно оно представлено. В результате обнаруживаемые на выборке закономерности могут оказаться не адекватными закономерностям, присущими генеральной совокупности. И от этого риска нельзя застраховаться ни человеку, ни каким бы то ни было формальным методам, даже если в их описании используются слова «независимость» и «представительность». Важно отметить, что человек, используя имеющиеся модели классов, готов осторожно адаптировать эти модели при появлении новых объектов. Человек включает новый объект в некоторый класс, если присутствие этого объекта гармонично согласуется с имеющейся моделью данного класса.

Отсюда вытекает требование **гармоничности** методов DM. Желательно, чтобы качество обучения (модели) оценивалось количественной мерой, которая должна меняться при включении нового объекта в состав того или иного класса. Максимального значения эта мера должна достигать при включении нового объекта в «свой» класс.

5. Все методы анализа данных используют ту или иную меру «близости» или «сходства». Человеческие способности оценивать отдаленное сходство или находить тонкие различия отличаются очень высокой эффективностью. Чем ближе формальная мера сходства будет **имитировать эти способности**, тем успешнее будут работать алгоритмы DM. Исследование разных мер сходства привело нас к выводу о целесообразности использования меры, которую мы назвали функцией конкурентного сходства или FRiS-функцией. Эта функция, как нам кажется, хорошо имитирует человеческий механизм оценки сходства. Оказалось, что на базе FRiS-функции удается построить эффективные алгоритмы решения всех основных задач распознавания образов [11, 12]. Эти алгоритмы удовлетворяют всем вышеперечисленным требованиям: они обладают свойствами совместимости и гармоничности, потенциальной оптимальности и инвариантны к соотношению числа объектов и числа признаков, характеру распределений и видам зависимостей между признаками. Поясним, что такое FRiS-функция.

5. Функция конкурентного сходства

Попробуем сформулировать свойства, которыми должна обладать функция сходства F, чтобы хорошо имитировать человеческий механизм оценки сходства и различия.

1. В литературе описаны десятки различных мер сходства [13]. Как правило, в этих мерах сходство контрольного объекта Z с эталонами носит **абсолютный** характер и зависит только от расстояний до этих эталонов. Но легко убедиться, что человеческое восприятие похожести носит **относительный** характер. Чтобы ответить на вопросы типа «Близко – далеко?», «Похож – не похож?» нужно знать ответ на вопрос «По сравнению с чем?». Отсюда следует, что функция F должна отражать **относительное значение** сходства в зависимости от особенностей конкурентного окружения.

Отметим, что этим свойством обладают некоторые существующие алгоритмы распознавания. Так, например, в правиле k ближайших соседей (*kNN*) решение о принадлежности объекта Z первому образу принимается не в том случае, когда расстояние r_1 до него «мало», а когда оно меньше расстояния r_2 до любого другого конкурирующего образа.

2. Человек может оценивать меру сходства не только в слабой шкале порядка («больше похож на А, чем на В»), но и давать **количественную меру** сходства («на сколько больше похож на А, чем на В»). Крайние значения функция F_i сходства объекта Z с эталоном i-го образа S_i должна принимать в двух случаях: +1, если объект Z совпадает с эталоном S_i, и −1, если объект Z совпадает с эталоном образа-конкурента S_j. При одинаковой похожести объекта на эталоны конкурентов функция $F_i = 0$.

Этими свойствами обладает функция

$$F_{i/j} = \left(r_j - r_i\right) \Big/ \left(r_i + r_j\right),$$

которую мы называем функцией конкурентного сходства или FRiS-функцией (от слов **F**unction of the **Ri**val **S**imilarity). Приведем примеры использования FRiS-функции для построения некоторых алгоритмов Data Mining.

6. Построение решающих правил (алгоритм FRiS-Stolp)

Для решения задачи типа D необходимо выбрать объекты-эталоны, с которыми будут сравниваться контрольные объекты. Выбор эталонов («столпов») для каждого образа можно осуществить с помощью алгоритма **FRiS-Stolp** [11]. Поясним его работу на примере распознавания двух образов.

Вначале выбираются столпы для первого образа S_1. Проверяется вариант, при котором первый случайно выбранный объект a_i является единственным столпом образа S_1, а в качестве столпов образа S_2 будем считать всех его представителей. Для всех объектов $a_j \neq a_i$ первого образа находится расстояние r_1 до своего столпа a_i и расстояние r_2 до ближайшего объекта второго образа. По этим расстояниям вычисляется значение FRiS-функции.

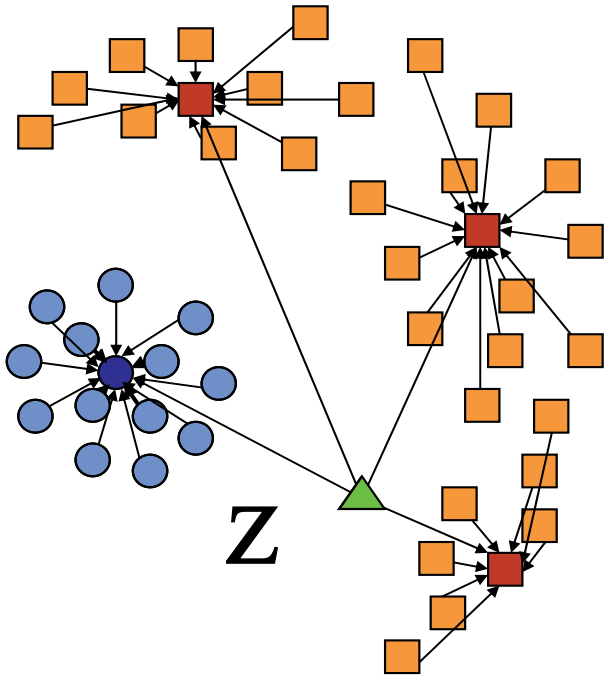

Рис.1. Распознавание принадлежности объекта Z к одному из двух образов, разделенных на кластеры

Находим те m_i объектов первого образа, значение функций сходства F которых выше заданного порога F^*, например, $F^*=0$. Величина m_i характеризует пригодность объекта a_i на роль столпа. Аналогичную процедуру повторяем, назначая на роль столпа все M_1 объектов первого образа по очереди. Находим объект a_i с максимальным значением m_i и объявляем его первым столпом A_{11} первого кластера C_{11} первого образа S_1.

Исключаем из первого образа m_i объектов, входящих в первый кластер. Для остальных объектов первого образа находим следующий столп повторением предыдущих шагов. Процесс останавливается, если все объекты первого образа оказались включенными в свои кластеры.

Восстанавливаем все объекты образа S_1 и для образа S_2 выполняем те же операции.

Итогом работы алгоритма FRiS-Stolp является решающее правило в виде списка эталонов (столпов), которые представляют каждый образ, и среднего значения функций сходства Fs объектов кластеров со своими столпами. Величина Fs служит количественной мерой качества обучения.

В алгоритме FRiS-Stolp первыми выбираются столпы, расположенные в центрах локальных сгустков и защищающие максимально возможное количество объектов с заданной надежностью. По этой причине при нормальных распределениях в первую очередь будут выбраны столпы, расположенные в точках математического ожидания. Это решение совпадает с оптимальным для этого случая, что отвечает требованию потенциальной оптимальности метода. Если распределения полимодальны и образы линейно не разделимы, столпы будут стоять в центрах мод. С ростом сложности распределения число столпов k будет увеличиваться. Следовательно, алгоритм инвариантен к виду распределения образов. Алгоритм работает при любом соотношении числа объектов и числа признаков.

Процесс распознавания с опорой на столпы очень прост и состоит в оценке конкурентного сходства контрольного объекта Z со всеми столпами и выбора образа S_i, чей столп в конкуренции со своим сильнейшим соперником S_t получил для объекта Z максимальное значение $F_{i/t}$ (Рис. 1.). Включение нового объекта в состав образа S_i подтверждает закономерности, присущие этому образу (например, значение среднего расстояния от объектов до своих столпов), что отвечает требованию гармоничности метода.

Еще одним важным преимуществом такого решающего правила является возможность использования значения $F_{i/t}$ в качестве оценки надежности принятого решения при распознавании конкретного объекта. На рис. 2 представлены результаты распознавания объектов контрольной выборки, получивших разные значения функции сходства F.

Как и ожидалось, при значениях F, близких к 0, вероятность ошибочного распознавания близка к 50%. С увеличением значения функции сходства F вероятность ошибки P быстро уменьшается.

7. Выбор информативных признаков (алгоритм FRiS-GRAD)

Для решения задачи типа X методом фильтрации может быть использован любой алгоритм выбора подмножества признаков, например, алгоритм GRAD [14]. Он позволяет автоматически указать как состав, так и наилучшее количество характеристик. Здесь мы обращаем внимание не на методы направленного перебора, а на новый критерий информативности, основанный на использовании FRiS-функции [15].

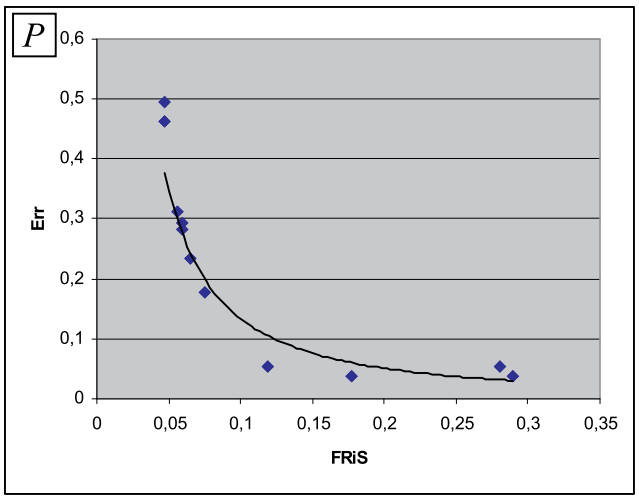

Рис. 2. Вероятность ошибки распознавания P в зависимости от величины функции сходства F

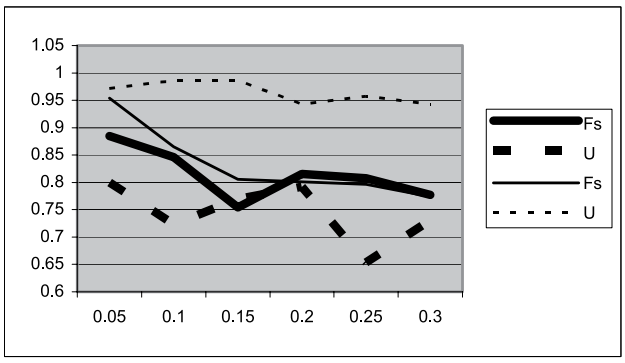

Рис. 3. Результаты обучения и распознавания по критериям U и F_s при разных уровнях шумов: тонкие линии – обучение, жирные – контроль

Как уже отмечалось, для оценки информативности признаков или их сочетаний часто используется прямой критерий в виде доли U правильно распознанных объектов обучающей выборки в режиме скользящего экзамена (OLO) или методом Cross-Validation (CV). Главным недостатком этого критерия является то, что он не учитывает надежность распознавания объектов, которые распознаны правильно, и грубость ошибки для тех объектов, которые распознаны неправильно. Нами было показано, что учесть эти особенности можно, если использовать в качестве косвенного критерия информативности среднее значение нормированной функции сходства F_s всех объектов обучающей выборки с эталонами своих образов.

Преимущества описанного критерия информативности F_s перед критерием U можно проиллюстрировать результатами их экспериментального сравнения. Исходные данные состояли из 200 объектов двух образов (по 100 объектов каждого образа) в 100-мерном пространстве. Признаки генерировались так, чтобы они обладали разной информативностью. В итоге около 30 признаков оказывались в той или иной степени информативными, а остальные признаки генерировались датчиком случайных чисел и были заведомо неинформативными. Дополнительно эта исходная таблица искажалась шумами разной интенсивности и при каждом уровне шума (от 0,05 до 0,3) алгоритмом GRAD выбирались наиболее информативные подсистемы размерности n в диапазоне от 1 до 22. При этом в режиме CV для обучения случайно выбиралось по 35 объектов каждого образа, а на контроль предъявлялись остальные 130 объектов. Результаты сравнения критериев F_s и U показаны на рис. 3.

Результаты контроля показывают, что критерий U дает завышенные оценки качества выбранных признаков. Косвенный критерий F_s обладает более высокими прогностическими свойствами и помехоустойчивостью по сравнению с прямым критерием U. Этот эффект можно объяснить тем, что критерий U реагирует на события, связанные только с фактом попадания или не попадания объекта в свою область пространства. А как далеко или близко объект находится от разделяющей границы значения не имеет. Оценку U формируют объекты, находящиеся в районе границы, т.е. те, что расположены в приграничных хвостах распределения и отражают редкие события. Критерий же F_s оценивает особенности распределения всех объектов. Свой вклад в оценку F_s вносят как те объекты, что расположены в хвостах распределений, так и те, которые находятся в районах локальных максимумов распределений. Ситуацию можно пояснить при помощи рис. 4.

Как в первом, так и во втором случае линейная решающая граница обеспечивает безошибочное распознавание двух образов. Критерий U не различает эти ситуации и дает оценки, равные 100%. В то же время критерий F_s считает случай 2 более предпочтительным. Для случая 1 он равен 0,59, а для случая 2 равен 0,71.

8. Оценка пригодности признаков (проблема А.Н. Колмогорова)

В 1933 году А.Н. Колмогоров [16] обратил внимание на задачи, в которых значительная часть характеристик играет роль случайного шума. Чем больше таких характеристик, тем выше вероятность обнаружения «псевдоинформативного» набора из шумовых предикторов. Вопрос А.Н. Колмогорова о том, как убедиться в неслучайности выбранных признаков, в их «пригодности» для дальнейшего использования, не теряет своей актуальности.

Мы предлагаем один вариант решения проблемы Колмогорова. Для сравнения результатов описанных экспериментов с чисто случайными результатами было создано 10 вариантов случайных таблиц такого же размера $M=200$, $N=100$. Два образа (по 100 объектов) были сформированы методом случайного выбора. По этим данным для каждой размерности подсистем n выбирались наиболее информативные признаки и определялись значения критерия F_s. Оказалось, что они лежат в «случайном коридоре» с границами от 0,61 до 0,67. Значения F_s для подсистем, найденных по исходной таблице, лежат значительно выше этого коридора (>0.85) и потому могут считаться неслучайными.

Опираясь на приведенные результаты, можно сформулировать следующую практическую рекомендацию. По обучающей таблице $N*M$ определя-

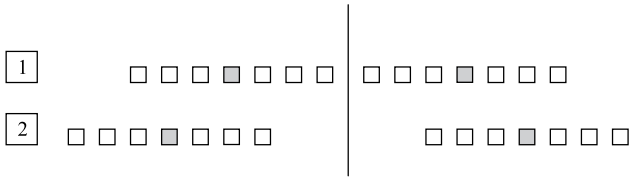

Рис. 4. Два случая распределения обучающих объектов двух образов.

ется значение F_s для наилучшей подсистемы из n^* признаков. Затем формируется серия случайных таблиц такого же размера N на M, и по ним находятся значения Fs для «лучших» подсистем той же размерности n^*. По расстоянию между значением критерия F_s подсистемы, выбранной в реальной таблице, и границами «случайного коридора» значений F_s, полученных на случайных таблицах того же размера, можно судить о степени неслучайности, пригодности выбранных подсистем.

9. Построение классификаций (алгоритм FRiS-Tax)

При решении задачи типа S автоматическая классификация объектов в виде иерархии классов или списка классов одного иерархического уровня может делатья с помощью алгоритма «FRiS-Tax» [17]. Его работа состоит из двух этапов. На первом этапе алгоритмом FRiS-Cluster выбираются объекты, находящиеся в центрах локальных сгустков объектов. Такие объекты становятся эталонами (столпами) кластеров. На втором этапе с помощью алгоритма FRiS-Class происходит процедура укрупнения кластеров в классы (таксоны) путем объединения некоторых соседних кластеров в один класс. Это позволяет создавать классы произвольной формы, не обязательно линейно разделимые.

Если найти среднее значение функции конкурентного сходства F_s всех объектов со столпами своих кластеров, то эта величина может характеризовать качество кластеризации. Оказалось, что при изменении количества кластеров k локальные экстремумы функции $F_s = f(k)$ имеют место при таких значениях k, которые эксперты считают наиболее предпочтительными. Это позволяет **автоматизировать выбор наилучшего количества кластеров** в заданном диапазоне значений k.

Делалось сравнение алгоритма FRiS-Tax с другими алгоритмами, оперирующими понятием центра кластера – с алгоритмом k-means [18,19] и Forel [2]. Результаты показали, что он превышает их по качеству получаемых решений (см. рис. 4).

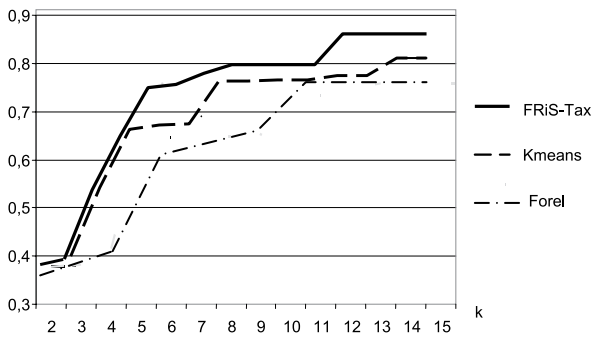

Рис. 5. Сравнение качества трех алгоритмов классификации

10. Применение FRiS-функции при решении других задач DM

Использование FRiS-функции позволяет легко реализовать алгоритмы для решения задач комбинированного типа [11]. Так, задача типа DS – одновременного построения классификации (S) и решающего правила (D) – решается прямо в процессе таксономии методом FRiS-Tax: таксоны описываются эталонами (столпами) кластеров, по которым в дальнейшем ведется распознавание новых объектов. Нами реализованы алгоритмы типа DX (построение решающего правила в наиболее информативном подпространстве признаков) и SX (таксономия в наиболее информативном подпространстве). Последний случай фактически совпадает с предельным по сложности алгоритмом комбинированного типа SDX. При этом признаки выбираются с помощью алгоритма FRiS-GRAD, а таксономия в каждом подпространстве делается алгоритмом FRiS-Tax, который попутно выдает и решающие правила в виде системы столпов.

Возможность применения FRiS-функции в алгоритмах заполнения пробелов (задача типа Z), прогнозирования (задача типа P) и поиска ассоциаций (задача типа A) так же не вызывает сомнений. Из вышесказанного можно сделать следующий вывод: **все задачи DM могут решаться методами, использующими общую основу в виде функции конкурентного сходства.**

11. Усиление эмпирической теории DM

Эмпирическая теория, описывающая общие свойства современных методов Data Mining, может быть представлена в следующем виде: $H_1 = <W_1, O_1, V_1, T_1>$, где

W_1 – множество всех возможных методов решения задач DM основных и комбинированных типов; O_1 – перечень характеристик, по совокупности которых эти методы отличаются друг от друга, и язык V_1 для записи значений измеряемых характеристик методов. Характеристики, которыми должен обладать любой метод G усиления эмпирических гипотез, таковы [1]:

1. Универсальность: метод G применим к любой допустимой паре $<h_0, pr_0>$. Здесь h_0 – исходная гипотеза, а pr_0 – обучающий протокол.

2. Нетривиальность: имеются допустимые пары, на которых метод G дает усиление гипотезы h_0, т.е. $G(h_0, pr_0) = h_1$, причем h_1 сильнее h_0, она точнее отражает действительность.

3. Последовательность: если гипотеза h_1 усилена за счет информации, содержащейся в протоколе pr_0, то она должна считать этот протокол допустимым.

Конкретные методы DM, обладающие указанными свойствами, могут отличаться друг от друга следующими характеристиками:

– типы задач, на которые ориентирован метод;

- требования к статистической обусловленности данных;
- ориентация на виды законов распределения;
- типы шкал, с которыми работает метод;
- метрика пространства признаков;
- критерий успешности решения задачи;
- трудоемкость метода;
- наличие положительного опыта применения метода;
- наличие сопровождающей информации (объяснений, инструкций) и т. д.

T_1 – тестовый алгоритм, который по значениям этих характеристик делит методы на допустимые и недопустимые.

Современное сообщество разработчиков и пользователей методов DM обращает основное внимание на три последние характеристики.

Описанные выше исследования функции конкурентного сходства показывают возможность единого подхода к решению разных задач, что обеспечивает совместимость методов при решении задач комбинированного типа. Кроме того, применение FRiS-функции для оценки качества предлагаемых вариантов решений позволяет повышать надежность принимаемых решений. Эти результаты позволяют дополнить перечень характеристик, которыми должен обладать допустимый метод DM, и внести изменения в тестовый алгоритм T, которые увеличивают его фильтрующие свойства.

Предлагаемое усиление общей эмпирической теории DM состоит в следующем. Представляющая ее теория H_2 отличается от H_1 тем, что в перечень характеристик метода добавляются следующие из описанных выше желательных характеристик:

- согласованность результатов при решении задач комбинированного типа;
- инвариантность метода к виду закона распределения;
- инвариантность к статистической обусловленности (соотношению числа объектов и признаков);
- гармоничность метода;
- потенциальная оптимальность метода.

Кроме того, к перечню значений характеристики «критерий успешности решения задачи» добавляется значение «косвенный критерий, основанный на конкурентной функции сходства F».

Тестовый алгоритм теории H_2 будет запрещать те методы решения задач комбинированного типа, в которых для разных частей задачи используются разные подходы; методы, зависящие от типа распределений, от соотношения числа объектов и признаков и методы, не обладающие свойством потенциальной оптимальности и гармоничности. Что же касается критерия проверки качества решений, то новый тестовый алгоритм более категоричен: $T(pr) = 1$, если используется FRiS-критерий, и $T(pr) = 0$ в противном случае.

Возможен и компромиссный вариант усиления теории H_1 до теории H_2*, тестовый алгоритм которой будет оценивать не все дополнительные характеристики метода, а лишь некоторую их часть. Например, он не будет требовать от метода инвариантности к законам распределения, и допускать использование традиционного метода оценки информативности по числу ошибок U. Такие варианты теорий будут занимать промежуточную позицию между теориями H_1 и H_2.

Потенциальная опровержимость Q_2 предложенной теории H_2 существенно выше Q_1 современной теории Data Mining. Проведенная серия экспериментов подтверждает правомочность описанного усиления потенциальной опровержимости эмпирической теории DM. Дальнейшее усиление таких характеристик теории, как степень подтвержденности P_2 и объясненности E_2 высокой эффективности нового подхода, будут продолжены.

12. Заключение

Знания о современных методах DM позволяют систематизировать их в виде онтологии и обобщить в форме эмпирической теории, которая отражает основные структурные элементы будущей теории DM. Для того, чтобы представленная эмпирическая теория стала соответствовать общепринятому понятию глубокой научной теории, требуется еще большая работа. Она связана с продолжением исследований методов индуктивного вывода, уточнением перечня характеристик методов DM и разработкой способов измерения их значений, разработкой вариантов тестовых алгоритмов, которые позволяли бы выбирать допустимые методы, адекватные конкретной задаче.

Для развития существующих методов DM предложена функция конкурентного сходства (FRiS-функция), которая имитирует человеческие способы оценки сходства и может использоваться в качестве универсальной основы для алгоритмов, решающих все основные и комбинированные задачи DM. Алгоритмы, основанные на FRiS-функции, применимы для решения задач с любой степенью обусловленности и при любом характере распределения анализируемых объектов в пространстве признаков. Использование FRiS-функции в качестве косвенного критерия для оценки качества обучения повышает точность этой оценки и позволяет решать такие новые задачи DM, как оценка пригодности признакового пространства, автоматическое определение числа кластеров, оценка надежности распознавании конкретного объекта при неизвестных законах распределений. Качество решений задач DM с помощью FRiS-функций не уступает качеству, получаемому другими известными методами.

Эти факты дают основание для усиления потенциальной опровержимости существующей эмпирической теории DM.

Благодарности

Работа была выполнена при поддержке РФФИ, гранты № 05-01-00241 и 08-01-00040. Автор выражает искреннюю благодарность К.Ф. Самохвалову, оказавшему большое влияние на развитие исследований в области методов эмпирического предсказания, и своим сотрудникам И.А. Борисовой, О.А. Кутненко и В.В. Дюбанову за активное участие в обсуждении представленной здесь проблемы и проведении большого количества машинных экспериментов.

Литература

1. Самохвалов К.Ф. О теории эмпирических предсказаний // Вычислительные системы, Вып. 55. Новосибирск, 1973, с. 3-35.
2. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Изд. ИМ СО РАН, Новосибирск, 1999. 273 с.
3. Загоруйко Н.Г., Самохвалов К.Ф., Свириденко Д.И. Логика эмпирических исследований. Изд. НГУ, Новосибирск, 1978. 68 с.
4. Н.Г. Загоруйко, Д.И. Свириденко. Формализация процесса углубления понимания // Эмпирические предсказания и распознавание образов. Новосибирск, 1967. Вып. 67: Вычислительные системы. С. 87-92.
5. Витяев Е.Е. Алгоритм естественной классификации // Анализ разнотипных данных (Вычислительные системы , 99). Новосибирск, 1983, с.44-50.
6. Borisova I., Zagoruiko N. Princips of natural classification // Proc. of 7-th Inter. Conf. on Pattern Recognition and Image Analysis: new information Technologies (PRIA-7-2004).St. Petersburg, 2004. pp.28-31.
7. Ivakhnenko A. G.: Polynomial theory of complex systems. IEEE Transactions on Systems, Man, and Cybernetics, SMC-1(1):364378, 1971.
8. Zagoruiko N.G., Gulyaevsky S.E., Kovalerchuk B.Ya. Ontology of Subject Domain «Data Mining» // Int. Journ. «Pattern Recognition and Image Analysis» 2007 (in press).
9. Журавлев Ю.И. Избранные научные труды. Изд. URSS, Москва, 1988. 420 с.
10. Лбов Г.С., Старцева Н.Г. Логические решающие функции и вопросы статистической устойчивости решений. Изд. ИМ СО РАН, Новосибирск, 1999. 212 с.
11. Борисова И.А., Дюбанов В.В., Загоруйко Н.Г., Кутненко О.А. Использование FRiS-функции для построения решающего правила и выбора признаков (задача комбинированного типа DX) // Труды Всероссийской Конференции «Знания-Онтологии-Теории» (ЗОНТ-07). Новосибирск, ИМ СО РАН, 2007 г. Том 1, сс. 37–44.
12. N.G. Zagoruiko, I.A. Borisova, V.V. Dyubanov, O.A. Kutnenko. Methods of Recognition Based on the Function of Rival Similarity // Pattern Recognition and Image Analisys. Vol 18, №1, 2008, pp. 1-6.
13. Воронин Ю.А. Начала теории сходства. Изд. ВЦ СО АН СССР, Новосибирск, 1989 г.,120 с.
14. Загоруйко Н.Г., Кутненко О.А. Алгоритм GRAD для выбора признаков // Труды VIII Межд. конференции «Применение многомерного статистического анализа в экономике и оценке качества». Изд. МЭСИ, Москва, 2006, сс.81-89.
15. . И.А. Борисова, Н.Г. Загоруйко, О.А. Кутненко. Критерии информативности и пригодности подмножества признаков, основанные на функции сходства // Заводская лаборатория. №1, том 74, 2008 г. С. 68-71.
16. Колмогоров А.Н. К вопросу о пригодности найденных статистическим путем формул прогноза. – Заводская лаборатория. 1933. №1. С. 164-167.
17. Борисова И.А. Алгоритм таксономии FRiS-Tax. Научные труды НГТУ, Новосибирск, 2007. Том 328 , С. 3-12
18. М.И. Шлезингер. О самопроизвольном разделении образов // Читающие автоматы и распознавание образов. Изд. «Наукова думка», Киев, 1965. СС. 46-61.
19. J. MacQueen. Some methods for classification and analysis of multivariate observations.// Proceedings of the 5th Berkley Symposium on Mathematical Statistic and Probability, Vol. 1, University of California Press, 1967, pp. 281-297.