ИСПОЛЬЗОВАНИЕ FRIS-ФУНКЦИЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ SDX

Ирина Борисова, Николай Загоруйко

***Аннотация**: Рассматривается задача структуризации избыточного набора информации, выявления основных закономерностей, содержащихся в нем с помощью аппарата FRiS-функций. В результате решения этой задачи (задачи SDX) на основе исходного множества объектов строится его сокращенное описание в терминах классов и существенных признаков. Данное описание снабжено системой правил, позволяющих восстанавливать значения всех признаков на основе существенных и находить место новым объектам в системе построенных классов.*

***Ключевые слова**: Распознавание образов, выбор признаков, натуральная классификация, функция конкурентного сходства.*

***ACM Classification Keywords**: I.5.2. Pattern Recognition*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

Введение

Формализация человеческой способности к анализу информации дает возможность частично наделять этой способностью искусственные объекты – компьютеры. Даже самые примитивные модели анализа данных, перенесенные на компьютеры, позволяют достигать значительных результатов, так как использование этих моделей позволяет машинам решать задачи, недоступные человеку из-за своей громоздкости и трудоемкости. Это становится особо актуально в последнее время, когда накопление информации в различных прикладных областях идет с огромной скоростью и ее обработка в принципе невозможна без помощи компьютера.

Одним из наиболее важных этапов обработки информации нам представляется ее систематизация и упрощение – представление в виде, доступном для понимания, более подробного исследования и дальнейшего использования. Человеком для этого используются различные приемы, многие из которых формализованы в рамках предметной области, называемой интеллектуальным анализом данных, и относятся к задачам распознавания образов. Вот основные из них :

1. **Сокращение числа рассматриваемых объектов**. Вместо изучения каждого отдельного представителя выборки рассматриваются классы сходных объектов. Похожесть в рамках класса позволяет заменять множества объектов из этого класса неким эталонным (идеальным) образцом, реализациями которого эти объекты являются.
2. **Упрощение описания классов.** Исходное описание класса в виде прямого перечисления всех объектов, попавших в него, заменяется описанием в виде обобщающего правила (логического решающего правила, линейной разделяющей границы и т.п.). Построение описаний уже существующих классов в виде решающих правил той или иной степени сложности, позволяет более четко представить структуру этих классов, их однородность.
3. **Сокращение числа учитываемых и используемых признаков**. Достигаться оно может как за счет исключения слабых, неинформативных, несущественных, случайных, шумящих признаков,

так и за счет выделения такой подсистемы информативных, существенных признаков, по которой можно восстановить все остальные неслучайные признаки с достаточной степенью точности.

В данной статье предпринимается попытка использовать формальные реализации когнитивных способностей человека для построения алгоритма решения одной из достаточно общих задач распознавания образов, когда перед исследователем оказывается набор данных единой природы (из ограниченной предметной области), представленный в виде таблицы «объект-свойство». При этом относительно представленного набора можно предположить лишь одно – он достаточно полно отражает многообразие объектов этой природы (предметной области) и многообразие признаков, их описывающих. Задачей же исследователя является структуризация этого возможно избыточного набора информации, представление его в виде, удобном для дальнейшего анализа и использования человеком.

Задача такого приведения исходной информации к виду, удобному для восприятия человеком, нами формулируется в терминологии задач комбинированного типа как задача типа **SDX** – задача одновременного формирования образов (задача **S**) с решающими правилами для их распознавания (задача **D**) в наиболее информативном подпространстве признаков (задача **X**).

Задачи основных типов такие как задача построения решающих правил, задача группировки объектов (таксономии), задача выбора системы информативных признаков хорошо известны и давно решаются в области распознавания образов. Однако при решении задач комбинированного типа для решения задач основных типов, из которых они состоят, целесообразно использовать единый подход, опирающийся на одну и ту же базовую гипотезу. В качестве единого базиса для решения различных задач распознавания образов мы используем метод оценки близости между объектами, основанный на функции конкурентного сходства (FRiS-функции).

Использование FRiS-функции позволило нам построить внутренне непротиворечивые и эффективные алгоритмы для решения таких задач комбинированного типа, как **DX** (распознавания с одновременным выбором информативной системы признаков), **SD** (таксономии с одновременным построением решающего правила) , **SX** (таксономия с одновременным выбором информативной системы признаков). Их описание содержатся в ранее опубликованных статьях [1, 2]. Теперь же рассмотрим, как функция конкурентного сходства может быть использована при решении задачи **SDX** (таксономии с одновременным построением решающего правила в пространстве наиболее информативных признаков).

Функция конкурентного сходства

Кратко напомним, что мы называем функцией конкурентного сходства, и какие предпосылки определяют ее эффективность при решении задач анализа данных.

Человек является самой совершенной из ныне существующих распознающих систем. Если мы хотим, чтобы наши алгоритмы хорошо имитировали человеческие способности решать задачи распознавания, то мы должны использовать ту же меру сходства, которую использует человек. Человек считает сходство категорией не абсолютной, а относительной, и оценивает меру сходства в зависимости от конкурентной ситуации. Для ответа на вопрос «На сколько сильно объект *a* похож на объект *b*?», нужно знать ответ на вопрос «По сравнению с чем?»

Для измерения в шкале отношений меры сходства объекта *z* с конкурирующими объектами *a* и *b* предлагается пользоваться следующими соотношениями:

$F_{a/b}=(r_b-r_a)/(r_a+r_b)$ для сходства Z с объектом *a*

и $F_{b/a}=(r_a-r_b)/(r_a+r_b)$ для сходства Z с объектом *b*.

Здесь r_a и r_b – расстояния от z до a и b, соответственно. Функцию F мы и называем функцией конкурентного сходства или FRiS-функцией (от слов **F**unction of **Ri**val **S**imilarity). $F_{a/b}$ принимает значение +1, если z и a неразличимы, -1, если z совпадает с b, и 0, если объект z равноудален от объектов a и b.

Формулировка и общая схема решения задачи SDX

Как было сказано выше, человек в силу особенностей своего восприятия предпочитает иметь дело не со всеми m объектами, а с небольшим числом k групп (кластеров) этих объектов, описанных небольшим набором информативных (существенных) признаков Y, выбранных из их исходного множества X. Чтобы быть практически полезной, такое сокращенное описание выборки должно содержать систему решающих правил, в соответствии с которыми каждый новый анализируемый объект может быть отнесен к той или иной группе. Помимо решающих правил сокращенное описание выборки должно содержать систему индуктивных правил, устанавливающих связь между подмножеством существенных признаков и всеми остальными признаками, не вошедшими в базис классификации. По таким правилам для каждого объекта, входящего в образ, по значениям его информативных признаков можно восстанавливать значения остальных признаков.

Это подход согласуется с принципами построения естественных классификаций [3], рассматриваемых рядом авторов, как способ объединения объектов в группы «на основании общих, присущих им свойств, определяющих множество других свойств этих объектов, как известных, так и еще не известных ». При этом «количество свойств объекта, поставленных в функциональную связь с его положением в системе, является максимальным»[4]. Возможность предсказывать значения признаков объектов по их месту в классификации мы будем называть предсказательной способностью классификации.

Рассмотрим вариант этой задачи, когда каждая группа объектов определяется своим типичным представителем (столпом). Новый объект относится к той группе, столп которой оказался ближайшим к этому объекту в пространстве информативных (существенных) характеристик. В качестве прогнозируемых значений признаков, не вошедших в число существенных, для этого объекта берется их значение для соответствующего столпа. Для оценки надежности такого рода прогноз мы используем функцию конкурентного сходства, которая измеряет близость между объектом и эталоном с учетом конкурентной ситуации.

В результате для фиксированного набора столпов $S \subseteq A$, где A-исходное множество объектов, и некоторого множества информативных признаков $Y \subseteq X$, где X - исходное множество признаков, определим качество, с которым выбранный набор данных $<S,Y>$ описывает исходный набор $<A,X>$ как:

$$Q_F(S,Y) = \sum_{a \in A} F_X(a, s^* \mid s^* = \arg\min_{s \in S} \rho_Y(a,s))$$

где F_X –функция конкурентного сходства в пространстве X, ρ_Y – метрика в пространстве Y. Задача же состоит в выборе такой пары $<S,Y>$, которая обеспечит максимум функционалу Q_F. Чтобы получить достаточно качественное решение этой сложной задачи мы разобьем ее на две более простые и перейдем к рассмотрению задачи двухуровневой оптимизации:

$$Q_F(Y) = \sum_{a \in A} F_X(a, s^* \mid s^* = \arg\min_{s \in S_Y} \rho_Y(a,s)) \to \max_{Y \subseteq X},$$

где $S_Y = \arg\max_{\substack{S \subseteq A, \\ |S| \le m^*}} \sum_{a \in A} F_Y(a, s^* \mid s^* = \arg\min_{s \in S} \rho_Y(a,s))$.

Набор столпов S_Y для фиксированной подсистемы признаков Y отыскивается с помощью алгоритма таксономии FRiS-Tax[2], который опирается на использование функций конкурентного сходства и в процессе работы строит набор столпов, обеспечивающий максимум среднего значения функции конкурентного сходства по выборке.

При переходе к решению задачи таксономии, мы опираемся на допущение, что в пространстве Y существенных (информативных) характеристик классы, обладающие реальными предсказательными свойствами, должны образовывать компактные сгустки, и, как следствие, отыскиваться с помощью некоторой таксономической процедуры. Выбор же самого пространства Y после определения алгоритма для вычисления $Q_F(Y)$ может осуществляться одной из известных процедур направленного поиска (алгоритмом AdDel, GRAD, СПА), либо локального спуска.

Таким образом, сложная задача SDX сводится к серии более простых, решение которых позволяет представлять исходную выборку объектов в виде, наиболее удобном для анализа пользователем, согласованно выделяя группы похожих объектов, решающее правило для отнесения новых объектов к выделенным группам и информативные (существенные) признаки, наиболее полно, описывающие выборку.

Проверка на реальных данных

Следующие эксперименты проводились, во-первых, для выяснения того, насколько точно восстанавливает информацию о выборке алгоритм FRiS-SDX, реализующий общую схему решения задачи SDX, описанную в предыдущем параграфе. Во-вторых, ставилась задача оценить, насколько отсутствие информации о выборке (отсутствие априорной информации о разбиении объектов на классы, об информативности описывающих признаков) ухудшает качество решения задач распознавания образов. Насколько оправданным в том или ином случае является переход от основных задач распознавания к комбинированным, и насколько он позволяет восстанавливать эту информацию и тем самым менять качество решения задач в зависимости от того, какова доля информативных признаков в выборке.

За основу была взята таблица, содержащая 64 мерные описания различных вариантов написания 10 арабских цифр. Мы предполагаем, что подобное разбиение является естественным, а практически все признаки – в той или иной степени информативными. Примеры объектов выборки приводятся на Рисунке 1.

Рис.1 Примеры объектов выборки, состоящей из различных вариантов написания арабских цифр.

Обучающая выборка A, сформированная на основе этой таблицы, содержала 100 объектов, тестовая выборка B – 655 объектов. Кроме того рассматривались «раздутые» варианты тех же выборок. Так в выборках A' и B' помимо исходных 64 признаков содержалось 64 клона этих признаков с наложенным на них Гауссовым шумом, а также 64 чисто шумовых признака, никак не связанных ни с целевым признаком, ни с исходными описывающими признаками. В итоге, каждый объект в этих выборках описывался уже 192 признаками и соотношение числа в той или иной степени информативных признаков к общему числу признаков было 1:2. По аналогии формировались выборки A'' и B''. Но в них уже было 1024 шумовых

признака (всего 1152 признака), и доля информативных признаков составляла 1:9. На этих выборках решались следующие типы задачи распознавания :

1. **Задача построения решающего правила** (задача **D)**. Эта задача соответствует случаю, когда известно как разбиение объектов обучающей выборки на классы, так и то, что среди описывающих признаков нет заведомо неинформативных, способных ухудшить качество распознавания. Для ее решения на обучающей выборке запускался алгоритм FRiS-Stolp [1], а эффективность построенного решающего правила оценивалась через качество распознавания обучающей - Q_{st}, и тестовой выборки - Q_{ts}.
2. **Задача таксономии** (задача **S**). Эта задача соответствует случаю, когда относительно выборки известно, что большая часть признаков информативны, однако информация о принадлежности объектов к классам недоступна. Она решалась с помощью алгоритма FRiS-Tax[2]. Качество таксономии оценивалась следующим образом. Каждому полученному таксону присваивалось имя класса, чьих представителей в нем оказывалось большинство, а затем на полученной выборке строилось решающее правило алгоритмом FRiS-Stolp. Чем выше при этом оказывалось качество распознавания по построенному правилу исходной обучающей выборки Q_{st} и качество распознавания тестовой выборки Q_{ts} тем более похожая на исходную естественную классификацию таксономия у нас получалась .
3. **Задача построения решающего правила с одновременным выбором информативных признаков** (задача **DX**). Эта задача возникает тогда, когда нет уверенности, что все признаки, вошедшие в таблицу-объект-свойство являются информативными, более того высока вероятность появления шумящих, неинформативных признаков, искажающих общую картину. Для ее решения использовался упрощенный вариант алгоритма FRiS-GRAD [1], в котором для направленного поиска системы признаков применялся алгоритм AdDel [5], а информативность каждой тестируемой системы признаков оценивалась через качество описания этой системы признаков системой столпов, построенных алгоритмом FRiS-Stolp. Этот алгоритм запускался на обучающей выборки, а затем полученным решающим правилом в пространстве информативных характеристик распознавалась контрольная выборка.
4. **Задача построения таксономии с одновременным выбором информативных признаков** (задача **SX**). В этом случае недоступной считается как информация об информативности признаков, так и информация о разбиении объектов обучающей выборки на классы. Для решения этой задачи мы использовали тот же алгоритм, что и для решения задачи SDX. Единственным отличием являлось то, что система столпов, которые в последствии могли использоваться как решающее правило, в нем не сохранялись. Качество полученной таксономии, как и в случае задачи **S**, оценивалось через надежность распознавания обучающей и контрольной выборки в выбранном подпространстве признаков в системе классов, сформированной в этой таксономии .
5. **Задача таксономии с одновременным построением решающего правила в пространстве наиболее информативных признаков** (задача **SDX**). Как и в предыдущей задаче, здесь отсутствующей считается информация как об информативности, так и о классовой принадлежности. Эта задача решалась алгоритмом FRiS-SDX, реализующим общую схему, описанную в данной статье. В результате его работы строилась некоторая классификация в пространстве информативных с точки зрения предсказательной способности характеристик. Параллельно строилось решающее правило для распознавания новых объектов. Чтобы оценить качество решения данной задачи, как задачи SDX, мы распознавали исходную обучающую и контрольную выборку относительно построенной классификации по построенному решающему правилу в пространстве информативных характеристик.

По сути две последние задачи в данном случае являются эквивалентными, так как используя алгоритм FRiS-Tax для построения таксономии мы автоматически строим решающее правило, разница лишь в методике оценки качества получаемых решений. В задаче **SX** решающее правило строится отдельно, а в задаче **SDX** для распознавания используется система столпов, полученных в процессе таксономии.

Результаты всех экспериментов, для выборок с различным уровнем шумов приводятся в Таблице 1.

Таблица 1.

Тип задачи	(*A*,*B*)		(*A'*,*B'*)		(*A''*,*B''*)	
	Q_{st}	Q_{ts}	Q_{st}	Q_{ts}	Q_{st}	Q_{ts}
D	0.96	0.82	0.94	0.80	0.72	0.49
DX	0.87	0.66	0.87	0.66	0.81	0.65
S	0.90	0.75	0.81	0.68	0.68	0.47
SX	0.85	0.68	0.83	0.69	0.54	0.36
SDX	0.85	0.68	0.8	0.69	0.54	0.37

Как и ожидалось, полученные результаты не дают возможности однозначно ответить на вопрос, следует ли от задач основных типов переходить к задачам комбинированным. Так в случае, когда доля информативных характеристик в выборке велика (пары (*A*,*B*) и (*A'*,*B'*)), выбор информативной подсистемы может ухудшить общее качество решения задачи **DX**. Таким образом, как и предполагалось, отказ от предположения об информативности описывающих признаков и ухудшает качество распознавания, в случае когда эта информации достоверна. Однако, с ростом числа шумящих характеристик в выборке такая процедура становится необходимой и оправданной, что подтверждает эксперимент по решению задачи **DX** на выборках *A''* и *B''*.

При построении таксономии наоборот, добавление процедуры выбора информативной системы признаков оказывается оправданной лишь при относительно небольшом уровне шумов в выборке (выборка *(A',B')*) и значительно ухудшается с их ростом.

Отказ от информации о классовой принадлежности объектов обучающей выборки также негативно сказывается на качестве получаемых решений, однако, в некоторых случаях это негативное влияние сглаживается на контрольной выборке, которая распознается лучше на более компактной системе классов, построенной в процессе таксономии. Именно поэтому результаты распознавания контрольной выборки в задаче **SX** для выборок (*A,B*) и (*A',B'*), оказываются даже лучше результатов решения задачи **DX** для них же.

Таким образом

Стоит отметить, что подобные результаты также объясняются спецификой конкретной задачи, в которой практически все исходные признаки, описывающие выборку, являются информативными и слабо коррелированными между собой, так как распознавание цифр по их частичному написанию представляется проблематичным. Потому их уменьшение автоматически ведет к потере качества.

Заключение

1. Показана возможность решения задачи комбинированного типа SDX одновременного выбора классификации S, решающего правила D и информативного подмножества X наблюдаемых объектов.

2. Для оценки предсказательной способности классификации при этом используется среднее значение функции F_X сходства объектов обучающей выборки с эталонами своих образов.

3. Экспериментально показано что информация о разбиении объектов на классы, получаемая в процессе решения задачи SDX, а также SX, достаточно хорошо согласуется с имеющейся естественной классификацией этих объектов. При этом удается сократить число признаков в описании классов.

4. Задачи комбинированного типа целесообразно решать в условиях отсутствия информации об обучающей выборке, при подозрении, что в описании содержатся неинформативные признаки, при отсутствии разбиения на классы.

Благодарности

Данная работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, Грант № 08-01-00040.

Библиография

1. N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov and O. A. Kutnenko. Methods of recognition based on the function of rival similarity. Pattern Recognition and Image Analysis. 2008. Vol. 18, No.1, pp. 1-6.

2. Борисова И.А. Алгоритм таксономии FRiS-Tax. Научный вестник НГТУ, 2007, №3(28), стр. 3-12.

3. Zagoruiko N., Borisova I. Principles of natural classification. Pattern Recognition and Image Analysis 2005 Vol.15, No1, pp.27-29.

4. Л.А. Субботин Классификация. Москва, 2001.

5. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: Изд. ИМ СО РАН, 1999.

Информация об авторах

Ирина Борисова – *Институт Математики СО РАН, пр. Коптюга, дом 4, Новосибирск, 630090, Россия; e-mail: biamia@mail.ru*

Николай Загоруйко *- Институт Математики СО РАН, пр. Коптюга, дом 4, Новосибирск, 630090, Россия; e-mail: zag@math.nsc.ru*