Сходство и компактность*

Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.
zag@math.nsc.ru
Институт математики им. С. Л. Соболева СО РАН, Новосибирский государственный университет

В статье вводится понятие функции конкурентного сходства (FRiS-функции), с помощью которой можно оценивать сходство между объектами и образами, получать количественные меры компактности образов и информативности признакового пространства. Описывается опыт использования предлагаемых мер сходства и компактности для решения задачи из области молекулярной биологии.

Мера сходства

Сходство $S(a,b)$ двух объектов a и b обычно оценивается величиной, которая зависит от расстояния $R(a,b)$ между этими объектами. Предполагается, что свойства расстояний — симметричность, рефлексивность, неравенство треугольника — проецируются и на меру сходства. Однако при распознавании образов нас интересует мера сходства с другими свойствами. Будем рассматривать сходство контрольного объекта z с объектами a и b, которые являются представителями (ближайшими объектами или эталонами) образов A и B, так что слова «сходство с объектом» будут означать то же, что и слова «сходство с образом». Для принятия решения о принадлежности контрольного объекта z к образу A недостаточно знать расстояние $R(z,a)$. Нужно знать также расстояние $R(z,b)$ и определить, что расстояние $R(z,a)$ является наименьшим из них. Следовательно, нужно иметь не абсолютную, а относительную меру сходства, величина которой зависит от расстояний до представителей конкурирующих образов. Если оценивается сходство между тремя объектами — a, b и c, то при оценке похожести объекта a на объект b должны учитываться расстояния $R(a,b)$ и $R(a,c)$, а при оценке похожести объекта b на объект a должны учитываться расстояния $R(b,a)$ и $R(b,c)$. Следовательно, относительная мера сходства $\bar{S}$ не обладает свойством симметричности: $\bar{S}(a,b) \neq \bar{S}(b,a)$. Не выполняется для этой меры и неравенство треугольника: сумма сходств между вершинами треугольника $\bar{S}(a,b)+\bar{S}(a,c)$ может быть как меньше, так и больше сходства $\bar{S}(b,c)$. Так что сходство, в отличие от расстояния, не образует *метрического* пространства. Относительная мера сходства, учитывающая конкурентную ситуацию, образует пространство, которое мы называем *конкурентным*.

Некоторые известные алгоритмы распознавания используют относительную меру сходства. Например, в методе k ближайших соседей (kNN) новый объект z распознается как объект образа A, если расстояние $R(z,A)$ до k ближайших объектов этого образа не только мало, но меньше, чем расстояние $R(z,B)$ до k ближайших объектов конкурирующего образа B. Оценка сходства в этом алгоритме делается в шкале порядка.

Более сложная мера сходства используется в алгоритме RELIEF [1]. Чтобы определить сходство объекта z с образом A в конкуренции с образом B используется величина

$$W_{A/B} = \frac{R(z,B) - R(z,A)}{R_{\max} - R_{\min}},$$

где $R_{\max}$ и $R_{\min}$ — максимальное и минимальное расстояния между всеми парами объектов. Сформулируем следующие требования, которым должна удовлетворять мера $F_{a/b}(z)$ сходства объекта z с объектом a в конкуренции с объектом b.

1. Мера сходства должна зависеть не от характера распределения всего множества объектов, а от особенностей распределения объектов в окрестности объекта z.

2. Если оценивается мера сходства объекта z с объектом a, и ближайшим соседом z является объект b, $b \neq a$, то при совпадении объектов z и a мера $F_{a/b}(z)$ должна иметь максимальное значение, равное $+1$, а при совпадении z с b — максимальное отрицательное значение, равное -1. Во всех остальных случаях мера конкурентного сходства принимает значения от -1 до $+1$.

3. При одинаковых расстояниях $R(z,a)$ и $R(z,b)$ объект z в равной степени будет похожим на объекты a и b,и меры сходства $F_{a/b}(z)$ и $F_{b/a}(z)$ должны быть равны 0.

Предлагаемая нами функция конкурентного сходства FRiS (Function of Rival Similarity) удовлетворяет всем этим требованиям [2]:

$$F_{a/b}(z) = \frac{R(z,b) - R(z,a)}{R(z,b) + R(z,a)}.$$

Выбор эталонов

Для выбора эталонных образцов (столпов), на основании сходства с которыми будет оцениваться конкурентное сходство контрольных объектов с образами, нами предлагается алгоритм FRiS-Stolp. Этот алгоритм выбирает эталоны следующим способом. Проверяется вариант, при котором первый

*Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00040, Международного фонда «Научный потенциал» и гранта АВЦП Рособразования, проект № 2.1.1/3235.

случайно выбранный объект $a_i,\ i = 1, \ldots, M_A$ образа A является единственным его столпом, а в качестве столпов образа B используются все его M_B объектов.

1. Для каждого объекта $a_j,\ j \neq i$, образа A находим расстояние r_{ji} до столпа a_i и расстояние r_{jb} до ближайшего к нему объекта b образа B. По этим расстояниям вычисляем значение функции сходства:

$$F_{ji/b} = \frac{r_{jb} - r_{ji}}{r_{jb} + r_{ji}}.$$

Чем больше эта величина, тем лучше объект a_i защищает объект a_j от включения его в состав образа B. Добавим полученную величину к счетчику C_i^1.

2. Повторив шаг 1 для всех $(M_A - 1)$ объектов $a_j,\ j \neq i$, получим в счетчике C_i^1 сумму оценок сходства объектов образа A с объектом a_i. Разделив эту сумму на $(M_A - 1)$, получим оценку F_i «обороноспособности» объекта a_i:

$$F_i^1 = \frac{C_i^1}{(M_A - 1)}.$$

3. Проделав шаги 1 и 2 для всех M_A объектов, мы получим оценки «обороноспособности» каждого из них. Теперь нужно проверить объект a_i на толерантность к объектам образа B. Для этого оценим сходство с a_i всех объектов $b_q,\ q = 1, \ldots, M_B$, образа B в предположении, что роль столпа этого образа будет играть объект b_s, который является ближайшим соседом объекта b_q.

4. Вычислим величину $F_{qs/i} = \frac{r_{qi} - r_{qs}}{r_{qi} + r_{qs}}$ сходства объекта b_q со своим столпом b_s в конкуренции со столпом a_i и добавим эту величину в счетчик C_i^2. Если эта величина положительна, то это повышает шансы объекта a_i стать столпом образа A. И наоборот. Повторив шаг 4 для всех объектов образа B, мы получим оценку F_i^2 толерантности объекта a_i по отношению к объектам образа B:

$$F_i^2 = \frac{C_i^2}{M_B}.$$

5. Если v — стоимость ошибки первого рода, а w — стоимость ошибки второго рода, то общую оценку F_i эффективности объекта a_i в качестве столпа образа A примем равной

$$F_i = \frac{vF_i^1 + wF_i^2}{v + w}.$$

Чем больше величина F_i^1, тем меньше будет ошибок первого рода (пропуск цели). Чем больше величина F_i^2, тем меньше будет ошибок второго рода (ложная тревога). Так что, их совместный учет должен отражать соотношение цен этих ошибок.

6. Повторяя шаги 4 и 5, мы получим такие оценки для для всех M_A объектов образа A. В качестве первого столпа образа A выбираем тот объект a_i, которому соответствует наибольшая величина F_i.

7. Затем выполним шаги 1–6 для объектов b_s образа B, $s = 1, \ldots, M_B$. Выбираем объект b_s, который получил наибольшую величину F_s, и объявляем его первым столпом образа B.

8. Теперь образы представлены не всеми объектами, а только своими столпами. В новых условиях выбор столпов может дать другой результат. Для проверки этого повторим шаги 1–7 с той разницей, что в качестве столпов конкурирующих образов будем использовать их столпы, выбранные на предыдущем этапе. Опыт показывает, что одной такой проверки оказывается достаточно.

9. Найдем объекты, сходство которых со своими столпами превышают заданный порог F^*, например, $F^* = 0$. Эти объекты образуют первые кластеры соответсвующих образов A и B.

10. Если не все M объектов вошли в эти кластеры, то для остальных объектов повторим шаги 1–9. При этом в качестве столпов конкурирующих образов будем использовать все их столпы, выбранные на предыдущих этапах. Шаг 10 повторяем до шага, после которого все объекты обучающей выборки оказываются включенными в свои кластеры. В итоге образы A и B будут представлены k_A и k_B столпами, соответственно.

Если количество образов $K > 2$, то задача сводится к предыдущей следующим способом. При выборе столпов последовательно для каждого образа (A) объекты всех остальных образов объединяются в один конкурирующий образ (B).

При нормальных распределениях в первую очередь будут выбраны столпы, расположенные в точках математического ожидания. Если распределения полимодальны и образы линейно неразделимы, столпы будут стоять в центрах мод. Количество столпов зависит от компактности образов.

Процесс распознавания с опорой на столпы состоит в оценке функций конкурентного сходства контрольного объекта z с двумя самыми близкими столпами разных образов. Решение принимается в пользу того образа, на столп которого контрольный объект похож больше всего.

Оценка компактности и цензурирование выборки

Практически все алгоритмы распознавания основаны на использовании гипотезы компактности. При определении компактности часто оперируют такими нечеткими понятиями, как «достаточно малое количество граничных точек», «не слишком вычурная граница» и т. д. Хотелось бы получить количественную меру компактности, значение ко-

торой было бы прямо связано с ожидаемой надежностью распознавания.

Одна из мер такого рода предложена в [3] и состоит в вычислении профиля компактности. Для каждого из M объектов a_i обучающей выборки все остальные $(M-1)$ объектов упорядочиваются по их расстоянию до a_i. При движении вдоль этих упорядоченных списков от первой позиции $j=1$ до последней $j=M-1$ в каждой порядковой позиции определяется количество объектов m_j, которые не принадлежат тому образу, которому принадлежит объект a_i. Величины $V_j = m_j/M,\ j=1,\ldots,M-1$, и формируют профиль компактности. Чем компактнее образы, тем для большего числа первых порядковых номеров профиля $j=1,\ldots,M-1$ выполнено $V_j=0$. Переход от профиля к количественной оценке компактности может делаться разными способами. В работе [3] описывается связь между профилем компактности и функционалом полного скользящего контроля, который является естественной количественной оценкой и компактности, и обобщающей способности для метода kNN.

Для получения количественной оценки компактности можно использовать описанную выше FRiS-функцию. Будем оценивать компактность образа A в задаче распознавания K образов. При выборе столпов образа A мы получили оценки F_i для всех M_A его объектов. Компактностью $\mathcal{F}_\mathcal{A}$ образа A будем считать среднее значение этих величин:

$$\mathcal{F}_\mathcal{A} = \frac{1}{M_A}\sum_{i=1}^{M_A} F_i.$$

Общая оценка $\mathcal{F}$ компактности K образов в данном признаковом пространстве, а следовательно, и информативности этого пространства, может быть получена путем арифметического или геометрического усреднения. Для минимизации ошибок всех образов в среднем следует использовать арифметическое усреднение:

$$\mathcal{F}' = \frac{1}{K}\sum_{j=1}^{K} \mathcal{F}_j.$$

Если нужно, чтобы компактность самого некомпактного образа была максимально возможной, тогда нужно использовать среднегеометрическую величину:

$$\mathcal{F} = \sqrt[K]{\prod_{j=1}^{K} \mathcal{F}_j}.$$

Наши эксперименты показывают, что критерий $\mathcal{F}$ обычно дает лучший результат по сравнению с критерием $\mathcal{F}'$. Описанная мера компактности тем больше, чем выше плотность объектов внутри образов, и чем дальше образы отстоят друг от друга. Она используется в качестве меры информативности признакового пространства в алгоритме FRiS-GRAD [2].

Найденные в процессе выбора столпов оценки F_i позволяют наметить пути решения проблемы «цензурирования» выборки. Оценка F_i у объекта, находящегося в центре локального сгустка своих объектов, будет больше, чем у периферийных объектов. Для объектов, оказавшихся в окружении чужих объектов, величина F_i может иметь отрицательное значение. Такой объект будет приводить к увеличению числа столпов и ухудшать качество распознавания. По этой причине эти объекты целесообразно исключить из дальнейшего рассмотрения. Процесс цензурирования состоит из последовательного исключения объектов и пересчета компактности оставшихся объектов. Сначала исключается объект, обладающий наименьшим значением величины F_i. После пересчета компактности обнаружится, что она увеличилась. Одновременно выявляется другой объект с минимальным значением F_i, который является кандидатом на очередное исключение. Если этот процесс не останавливать, то максимальная компактность, равная 1, будет достигнута, когда останутся только объекты-столпы. Цензурирование должно остановиться на шаге, при котором достигает максимума критерий Q, отражающий два противоречивых желания: добиться максимального значения компактности при минимальном сокращении количества объектов обучающей выборки:

$$Q = f(\mathcal{F}, N_c/N_0),$$

где N_c/N_0 — доля выборки, сохранившейся после цензурирования. В настоящее время исследуются некоторые варианты этого критерия.

Ниже приводится пример использования описанных алгоритмов при решении одной из реальных задач.

Диагностика рака простаты по масс-спектрам белков

Анализируются данные о масс-спектре белковых форм, полученные с помощью спектрометра типа SELDI-MS-TOF [4]. Количество признаков (спектральных полос) — 15153. Представлены 4 класса пациентов с разным уровнем индекса PSA, характеризующего степень развития рака простаты: 63 здоровых пациента класса B имеют PSA < 1 ng/mL, 26 пациентов класса C имеют PSA $= 4 \div 10$ ng/mL, 43 пациента класса D имеют PSA > 10 ng/mL и 190 пациентов класса A имеют PSA > 4 ng/mL. Малое количество пациентов не позволяет разделить выборку на обучающую и контрольную. По этой причине будем обучаться на двух классах, а на контроль предъявлять объекты третьего класса.

О качестве обучения и распознавания будем судить, исходя из следующих соображений. Если упорядочить классы пациентов по степени проявле-

Таблица 1. Результаты экспериментов.

Обучение	Контроль	B	C	D
$[B\_D]$	A_{190} (> 4)	3		187
$[B\_D]$	C_{26} $(4 \div 10)$	0		26
$[B\_C]$	A_{190} (> 4)	1	189	
$[B\_C]$	D_{43} (> 10)	3	40	
$[C\_D]$	A_{190} (> 4)		168	22
$[C\_D]$	B_{63} (< 1)		49	14
$[B\_C\_D]$	A_{190} (> 4)	19	137	34

ния симптомов рака от самого здорового до самого больного, то класс B должен находиться в начале списка, за ним должен следовать класс C, и затем — класс D. Пациенты класса A должны оказаться среди пациентов классов C и D. Если построить правила для распознавания класса здоровых пациентов B от любого класса больных (например, класса C), то пациенты других классов больных (A и D) должны быть больше похожими на класс C, чем на B. Перебирая разные составы конкурирующих классов и фиксируя выбираемые при этом информативные характеристики, можно выделить подмножество характеристик, по которым классы будут отличаться друг от друга.

На первом этапе были сформированы две группы классов: первую группу представлял класс здоровых пациентов B, а во вторую группу были включены три класса больных пациентов — классы A, C и D. С помощью алгоритма FRiS-GRAD [2] в режиме Cross-Validation (10 этапов по 10% выборки на контроль) из 15153 признаков в состав 10 решающих правил было включено 24 признака. По этим правилам правильно распознано 275 объектов из 322 (85,4%). Надежность распознавания здоровых пациентов была равна 43 из 63 (68,3%), а больных — 232 из 259 (89,6%).

На следующем этапе делалась попытка из 24 найденных признаков выбрать информативные подсистемы для распознавания всех классов друг от друга. Результаты решения некоторых из этих задач представлены в таблице 1, из которой видно, что класс здоровых хорошо отличается от всех классов больных пациентов. Два класса больных (классы C и D) различаются хуже.

Кроме этой задачи описанными методами были успешно решены и другие задачи из области медицины. В задаче распознавания типов лейкемии по экспрессии генов, которая решалась многими другими авторами, получены результаты, превышающие результаты этих авторов [5]. В области физики успешно решена задача распознавания классов мелкодисперсных веществ по рентгеновским спектрам [6]. Использование FRiS-функции в задачах кластеризации и таксономии позволяет строить линейно неразделимые таксоны с автоматическим выбором наилучшего числа таксонов [2]. Достаточно успешным оказалось применение FRiS-функции в алгоритме прогнозирования. В международном конкурсе Data Mining CUP 2009 [7] по решению задачи прогнозирования спроса на разные книги в разных магазинах, участвовало 52 команды из Германии, США, Великобритании, Китая и других стран. Лучший результат получил оценку 17260, худший — 1938612. Наш результат 18353 оказался на четвертом месте. Общее свойство этих задач состояло в том, что количество признаков N на порядки превышало количество объектов M.

Выводы

Рассмотрение относительной меры сходства, учитывающей конкурентную обстановку, позволяет строить эффективные алгоритмы решения всех основных задач Data Mining. Функция конкурентного сходства дает возможность вычислять количественную оценку компактности образов и информативности признакового пространства и строить легко интерпретируемые решающие правила. Метод инвариантен к количеству образов, характеру их распределений и обусловленности обучающей выборки (соотношению между M и N). Трудоемкость метода позволяет использовать его для достаточно сложных реальных задач.

Литература

[1] *Kira K., Rendell L.* The Feature Selection Problem: Traditional Methods and a New Algorithm // Proc. 10th Nat'l Conf. Artificial Intelligence (AAAI-92). — 1992. — Pp. 129–134.

[2] *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* Methods of Recognition Based on the Function of Rival Similarity // Pattern Recognition and Image Analysis. — 2008. — V. 18. — Pp. 1–6.

[3] *Воронцов К. В., Колосков А. О.* Профили компактности и выделение опорных объектов в метрических алгоритмах классификации // Искусственный Интеллект. — 2006. — С. 30—33.

[4] *Ziener C., Foster P. S., Divall E. J., Hooker C. J., Langley A. J., Neely D.* Time-of-Flight corroboration on «conventional» ultra high intensity measurement // Central Laser Facility Annual Report. Chilton, UK. — 2001/2002.

[5] *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* Attributr selection through decision rules construction (algorithm FRiS-GRAD) // Proc. of 9th Intern Conf. Pattern Recognition and Image Analysis: New Information Technologies, Nizhni Novgorod, — 2008. — V. 2. — Pp. 335–338.

[6] *Богданов А. Б., Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А., Кучкин А. В., Мещеряков М. А., Миловзоров Н. Г.* Интеллектуальный анализ спектральных данных // Автометрия. — 2009. № 1. — С. 92–101.

[7] http://www.prudsys.de/Service/Downloads/bin/DMC2009_Ergebnisliste.pdf