УДК 519.95

Количественная мера компактности и сходства в конкурентном пространстве

Загоруйко Н.Г., Борисова И.А., Дюбанов В.В., Кутненко О.А.

Аннотация. Описываются меры сходства между объектами в метрическом и конкурентном пространствах. В качестве меры сходства, используемой в задачах классификации и распознавания образов, предлагается функция конкурентного сходства (FRiS). Эта функция позволяет создавать эффективные алгоритмы решения всех основных задач Data Mining, получать количественную оценку компактности образов и информативности признакового пространства и строить легко интерпретируемые решающие правила. Метод применим к задачам с произвольным количеством образов, любому характеру их распределений и обусловленности обучающей выборки (соотношению между M и N).

Ключевые слова: мера сходства, распознавание образов, компактность, информативность.

1. Введение

При формулировке эмпирических закономерностей часто используется понятие «сходство». Мера сходства играет ключевую роль при формировании классификации изучаемого множества объектов (кластеризации) и при распознавании принадлежности объектов к тому или иному классу. Специфика этих задач состоит в том, что мера сходства здесь является величиной относительной, она зависит не только от сходства объекта с определенным классом, но и от его сходства с другими (конкурирующими) классами. Такая мера, адекватная конкурентной ситуации, не удовлетворяет всем свойствам метрического пространства. В результате задача распознавания решается в пространстве, которое мы называем конкурентным, а мера сходства в этом пространстве называется функцией конкурентного сходства (FRiS, от слов **F**unction of **Ri**val **S**imilarity).

Использование FRiS-функции позволяет получать количественную оценку компактности образов, а величину компактности можно использовать в качестве критерия информативности признакового пространства. FRiS позволяет выбирать эталонные объекты (столпы) и строить эффективные решающие правила.

Полезность использования FRiS показывается на примере решения задач распознавания и прогнозирования из области генетической медицины и коммерции.

2. Мера сходства в метрическом и конкурентном пространствах

Сходство $S(a,b)$ двух объектов a и b в метрическом пространстве обычно оценивается величиной, которая зависит от расстояния $R(a,b)$ между этими объектами. Если максимальное парное расстояние между объектами множества (диаметр множества) принять за 1, то $S(a,b)=1-R(a,b)$. В литературе описано большое количество разных мер сходства такого рода [1 , 2].

Обычно предполагается, что свойства, которыми обладает расстояние, проецируются на свойства, которыми должна обладать мера сходства. Так, если расстояния между объектами находятся в диапазоне $1 \geq R(a,b) \geq 0$, то неотрицательность расстояния между объектами влечет неотрицательность их сходства: $1-R(a,b)=S(a,b)\geq 0$. Из симметричности расстояния $R(a,b)=R(b,a)$ следует симметричность сходства: $S(a,b)=S(b,a)$. При этом сходство объекта a с объектом b не зависит от сходства объекта a с другими объектами.

Это означает, что расстояние и сходство в метрическом пространстве рассматривается в качестве *абсолютных* категорий. Действительно, если известны используемые стандартные единицы (метры, микрометры и т.д.), то расстояние измеряется в абсолютной шкале. Каждый акт измерения в абсолютной шкале состоит в сравнении измеряемого объекта с эталонным объектом (эталоном веса, длины, времени и т.д).

Но меру сходства, используемую в распознавании образов, так измерять нельзя. Эталона свойства «сходство» не существует. При распознавании принадлежности объекта z к одному из двух образов A или B важно знать не только его расстояние до образа A, но и расстояние до конкурирующего образа B, и сравнивать эти расстояния друг с другом.

Следовательно, сходство в распознавании образов является категорией не абсолютной, а *относительной.* Чтобы ответить на вопрос «На сколько сильно z похож на a?», нужно знать ответ на вопрос «По сравнению с чем?». Адекватная мера сходства должна определять **относительную величину сходства,** зависящую от особенностей конкурентного окружения.

Все статистические алгоритмы распознавания учитывают конкуренцию между классами. Если в точке z априорная вероятность класса A равна P_A, а класса B – P_B, то для принятия решения эти величины сравниваются между собой, и решение принимается в пользу, например, класса А не потому, что P_A равна некоторой стандартной единице плотности или превышает определенный порог, а потому что $P_A > P_B$.

Для случаев, когда законы распределения образов не известны или когда количество признаков на порядки превышает количество объектов обучающей выборки, оперировать плотностями вероятностей нельзя. Обычно используются расстояния $R(z,a_i)$ между контрольным объектом z и эталонами (прецедентами) образов A_i, $i=1,2...$ K (K – количество образов) и решение принимается в пользу того образа, расстояние до эталона которого меньше, чем расстояния до эталонов других образов. Например, в методе «k ближайших соседей» (kNN) [3] новый объект z распознается как объект образа A, если расстояние R_A до этого образа меньше, чем расстояние R_B до ближайшего конкурирующего образа B. Оценка сходства в этом алгоритме делается в шкале порядка.

Более сложная мера сходства используется в алгоритме RELIEF [4]. Чтобы определить сходство объекта z с образом A в конкуренции с образом B используется величина, которая учитывает разницу в расстояниях до конкурентов в явном виде:

$$W_{zA/B} = (R_B - R_A) / (R_{max} - R_{min}), \qquad (1)$$

Нормализация разности расстояний по величине $(R_{max}-R_{min})$ представляется не удачной. Мера при этом зависит от общих свойств обучающей выборки, но не учитывает локальных особенностей распределения объектов в непосредственной близости от объекта z. Предельная величина сходства не ограничена. Если дисперсия парных расстояний между объектами обучающей выборки мала, то сходство может оказаться очень большим, вплоть до бесконечности.

Мы формулируем следующие требования, которым должна удовлетворять мера F сходства объекта z с объектом a:

1. Мера сходства должна зависеть не от характера распределения всего множества объектов, а от особенностей распределения объектов в непосредственной близости от 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 $F_{za/b}$ должна иметь максимальное значение, равное +1, а при совпадении z с b – максимально отрицательное значение, максимальную непохожесть на a, равную -1.

3. Во всех остальных случаях мера конкурентного сходства должна принимать значения между +1 и -1 и иметь вид непрерывной невозрастающей функции.

4. При одинаковых расстояниях R_a и R_b объект z в равной степени будет похожим или не похожим на объекты a и b и функции сходства $F_{za/b}$ и $F_{zb/a}$ должны быть равны 0.

Этим требованиям удовлетворяет любая сигмоидная функция. Мы предлагаем использовать следующий простой вариант функция конкурентного сходства FRiS:

$$F_{za/b}=(R_b-R_a)/(R_a+R_b) \qquad (2)$$

Как расстояние R между объектами, так и сходство F между ними не зависят от положения начала координат, поворота координатных осей и одновременного умножения их значений на одну и ту же величину. Следовательно, конкурентное сходство измеряется в достаточно сильной измерительной шкале – шкале отношений. Но независимые изменения масштабов разных координат меняют вклад, вносимый отдельными характеристиками в оценку и расстояния и сходства. Так что, сходство между объектами зависит от того, с какими весами мы учитываем характеристики при его оценке. Меняя веса характеристик, можно подчеркнуть сходство или различие между заданными объектами или их подмножествами, что обычно и делается при выборе информативных признаков и построении решающих правил в распознавании образов.

В отличие от расстояний мера сходства не удовлетворяет некоторым аксиомам метрического пространства. В частности, нарушается аксиома треугольника: сумма сходств между объектами, расположенными в вершинах треугольника $F_{a,b/c}+F_{b,c/a}$ может быть как меньше, так и больше сходства $F_{c,a/b}$. Это можно легко видеть на примере прямоугольного треугольника $<a,b,c>$, гипотенуза которого $<b,c>$ равна диаметру множества объектов, так что, длина гипотенузы равна 1. Пусть при этом катет $<a,b>$ имеет длину 0,8, а катет $<a,c>$ - длину 0,6. Тогда конкурентные сходства между вершинами треугольника (при обходе вершин по маршруту a,b,c) будут равны: $F_{a,b/c}$=(0.6-0.8)/1.4= - 0.107, $F_{b,c/a}$=(0.8-1)/1.8= - 0.111, $F_{c,a/b}$=(1-0.6)/1.6=0.250 . Сумма первых двух сходств равна - 0.218, что меньше третьего сходства 0.250. Не выполняется на сходствах и аксиома симметричности: $F_{c,a/b}$=0.250, а $F_{a,c/b}$ = 0.107. Наконец, величина сходства может быть как положительной, так и отрицательной. Формально значения сходства можно ограничить пределом от +1 до 0, чтобы выполнялась аксиома положительности значений. При этом нейтральный вариант сходства был бы равен 0.5. Однако, такая шкала значений представляется менее удобной для интерпретации результатов анализа. Наличие как положительных, так и отрицательных значений упрощает интерпретацию сходств и различий. По мере увеличения расстояния R между объектами z и a можно говорить вначале о большом сходстве объекта z с объектом a, затем об умеренном их сходстве, о наступлении одинакового сходства как с объектом a, так и b, об умеренном и затем большом несходстве с a, т.е. об отличии z от a. Совпадение объекта z с объектом b дает значение $F_{za/b}$ =-1, что означает максимальное отличие z от a.

Но и при сохранении свойства положительности можно видеть, что сходство, в отличие от расстояния, *не образует метрического пространства.* Отличие меры конкурентного сходства F от абсолютного сходства S в метрическом пространстве дает основание для того, чтобы пространство, образуемое этой мерой, называть пространством конкурентного сходства или *конкурентным пространством.*

Абсолютная мера сходства ($S=1-R$) давала трудно интерпретируемый ответ на вопрос: «Чему равно сходство объекта z с эталоном образа A?». Сходство в шкале порядка, используемое в методе kNN, отвечает на вопрос: «На эталон какого образа объект z похож больше всего?». Конкурентное сходство в шкале отношений, измеряемое с помощью FRiS-функции, отвечает на такой вопрос: «Чему равно сходство объекта z с эталоном образа A в соревновании с самым сильным конкурентом – эталоном образа B?».

3. Конкурентное сходство с учетом локальной плотности распределений

Для того, чтобы учесть различия в распределениях образов A и B мы используем λ-расстояния [6] от объекта z до ближайших объектов конкурирующих образов. Пусть распознавание ведется по сходству объекта z с ближайшими к нему объектами a_i и b_i конкурирующих образов A и B. Пусть расстояние между объектом z и a_i равно R_A. Расстояние от a_i до ближайшего соседа своего образа a_j, $a_j \neq a_i$ равно α. Тогда λ-расстояние от z до образа A будет равно $\lambda_A = R_A / \alpha$. Аналогичным способом можно получить λ-расстояние от z до образа B: $\lambda_B = R_B / \beta$. Тогда сходство объекта z с образом A, учитывающее локальный характер плотности распределения конкурирующих образов, будет зависеть от этих расстояний так: $F_{zA/B}=(\lambda_B - \lambda_A)/(\lambda_A + \lambda_B)$. Объект z распознается в качестве объекта образа A, если выполняется условие $F_{zA/B} \geq 0$. На рис. 1 показаны функции конкурентного сходства для одинаковых и разных плотностей распределения образов.

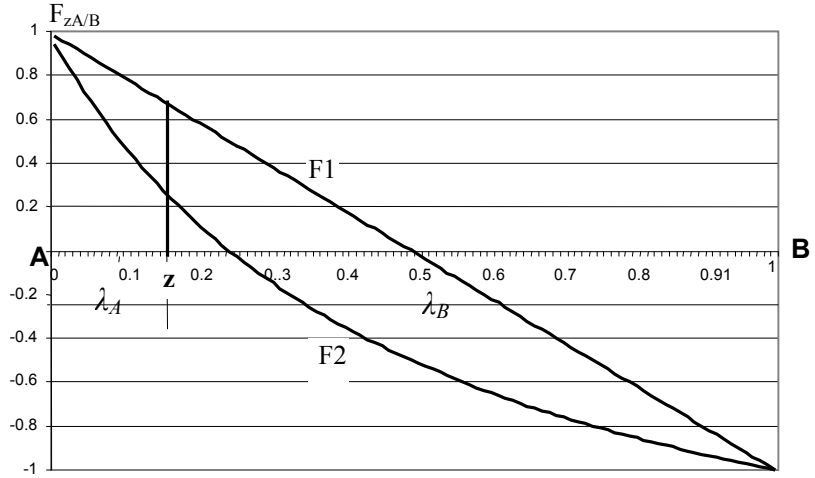

Рис. 1. Варианты функций конкурентного сходства объектов с эталоном образа A в конкуренции с образом B. Функции используют λ-расстояния и построены для случаев, когда плотности образов одинаковы (F1), и когда плотность образа A в три раза больше плотности образа B.

Граница между образами проходит через точку, в которой $F_{zA/B}=0$. Если плотности образов одинаковы, то это точка X=0.5, в которой расстояния до эталонов конкурирующих образов одинаковы. При разных плотностях граница смещается в сторону более плотного образа, что соответствует требованиям теории статистических решений. Для увеличения надежности оценок вместо расстояния от a_i до одного ближайшего соседа можно использовать среднее значение расстояний от объекта a_i до k своих ближайших соседей, как это делается в методе kNN, либо до всех объектов образа.

4. Построение решающего правила

Для распознавания образов необходимо выбрать объекты-эталоны, с которыми будут сравниваться контрольные объекты. Выбор эталонов («столпов») для каждого образа можно осуществить с помощью алгоритма **FRiS-Stolp**. В качестве столпов выбираются такие объекты, которые обладают высокими значениями двух свойств: обороноспособностью по отношению к объектам своего образа и толерантностью по отношению к объектам других образов. В результате выбираются такие столпы, на которые свои объекты похожи больше, чем на столпы конкурирующих образов.

Этот алгоритм работает при любом соотношении количества объектов M к количеству признаков N и при любом виде распределения образов. Он выбирает правила для

произвольного количества образов, но объяснять его работу будем на примере распознавания двух образов – *A* и *B*, количество объектов в которых M_A и M_B (см. рис.2).

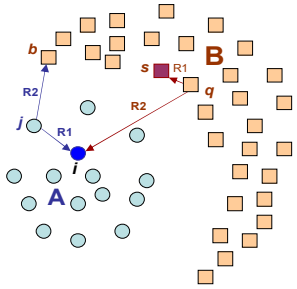

Рис. 2. К выбору эталонных объектов (столпов)

0. Признаковое пространство задано. Проверяется вариант, при котором первый случайно выбранный объект a_i, $i=1—M_A$, образа А является единственным его столпом, а в качестве столпов образа *B* используются все его M_B объектов.

1. От каждого объекта a_j, $i \neq j$, образа *A* находим расстояние *R1* до столпа a_i и расстояние *R2* до ближайшего объекта *b* образа В. По этим расстояниям вычисляем значение функции сходства $F_{ji/b} = (R2-R1)/(R2+R1)$. 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.

2. Повторив п.1 для (M_A-1) объектов a_j, $i \neq j$, получим в счетчике C'_i сумму оценок сходства всех объектов образа *A* с объектом a_i. Разделив эту сумму на количество (M_A-1), получим оценку F'_i обороноспособности объекта a_i: $F_i^{'} = C_i^{'}/(M_A - 1)$

3. Проделав процедуры пунктов 0, 1 и 2 для всех M_A объектов, мы получим оценки обороноспособности каждого из них. Чем выше обороноспособность эталона, тем меньше вероятность появления ошибок первого рода («пропуск цели»).

Теперь нужно проверить объект a_i на толерантность к объектам образа В. Чем она выше, тем меньше вероятность появления ошибок второго рода («ложная тревога»). Для этого оценим сходство с a_i всех объектов b_q, $q=1...M_B$, образа *B* в предположении, что роль столпа образа *B* будет играть объект b_s, который является ближайшим соседом объекта b_q. Примем, что R1- расстояние от b_q до b_s, а *R2* - расстояние от b_q до a_i.

4. Вычислим величину $F_{qs/i} = (R2-R1)/(R2+R1)$ сходства объекта b_q со своим столпом b_s в конкуренции со столпом a_i и добавим эту величину в счетчик C''_i. Если эта величина положительна, то объект a_i обладает толерантностью по отношению к объекту b_q, и добавление $F_{qs/i}$ к C''_i повышает шансы объекта a_i стать столпом образа *A*. И наоборот, отрицательное значение толерантности уменьшает эти шансы.

5. Повторив п. 4 для всех объектов образа *B* мы получим оценку F''_i толерантности объекта a_i по отношению к объектам образа *B*: $F_i^{''} = C_i^{''} / M_B$. Общую оценку F_i объекта a_i в качестве столпа примем равной $F_i = (F_i^{'} + F_i^{''})/2$.

6. Выполнив пп. 4 и 5 для всех M_A объектов, мы получим такие оценки для всех объектов образа *A*.

7. В качестве первого столпа образа *A* выбираем тот объект a_i, который получил наибольшую величину F_i.

8. Затем выполним процедуры пп. 0-7 для объектов b_s образа *B*, $s=1—M_B$. Выберем объект b_s, который получил наибольшую величину F_s, и объявляем его первым столпом образа *B*.

10. Первые столпы a_i и b_s были выбраны в условиях, когда им противостояли все объекты конкурирующих образов. Теперь образы могут быть представлены не всеми объектами, а только своими столпами. В новых условиях роль столпов может перейти к другим объектам. Для проверки этого повторим пп. 0-9 с той разницей, что в качестве столпов конкурирующего образа будем использовать его первый столп, выбранный на предыдущем этапе. Опыт показывает, что одной такой проверки оказывается достаточно: список первых столпов при дальнейших повторах практически не меняется. Решение относительно выбора первых столпов, полученное на данном этапе, принимаем в качестве окончательного.

11. Найдем объекты, сходство которых со своими столпами превышают заданный порог F^*, например, $F^*=0$. Если считать плотности распределений образов одинаковыми, то при расстоянии D между первыми столпами этому условию удовлетворяют все объекты, которые находятся от своих столпов на расстоянии R<D/2. Будем считать такие объекты принадлежащими первым кластерам образов *A и B*.

12. Если не все *M* объектов вошли в эти кластеры, то для остальных объектов повторим процедуры по п. 0-11.

13. П. 12 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).

Процесс распознавания с опорой на столпы очень прост, и состоит в оценке функций конкурентного сходства контрольного объекта *z* с двумя самыми близкими столпами разных образов. Объект *z* относится к тому образу, сходство со столпом которого максимально.

Можно применять два варианта стратегии распознавания – «без порога» и «с порогом». При первой стратегии решение о принадлежности объекта *z* принимается в пользу того образа, сходство с которым максимально, вне зависимости от того, на каком расстоянии от распознаваемых образов находится объект *z*. При второй стратегии решение принимается с учетом расстояния до ближайшего из конкурирующих образов. Если оно превышает некоторый порог, например больше расстояния *d* между ближайшими к объекту *z* столпами этих двух образов, то считается, что объект принадлежит новому (K+1)-му образу. Применять эту стратегию целесообразно, чтобы, например, при распознавании по весу двух классов мелких грызунов не отнести к одному из этих классов медведя.

Отметим некоторые особенности алгоритма FRiS-Stolp. Вне зависимости от вида распределения столпами выбираются объекты, расположенные в центрах локальных сгустков и защищающие максимально возможное количество объектов с заданной надежностью. При нормальных распределениях столпами в первую очередь будут выбраны объекты, ближайшие к точкам математического ожидания. Следовательно, при приближении к нормальным распределениям решение задачи построения решающих функций стремится к оптимальному. Если распределения полимодальны и образы линейно неразделимы, столпы будут стоять в центрах мод. Количество столпов зависит от компактности образов.

5. Компактность и информативность в конкурентном пространстве.

Практически все алгоритмы распознавания основаны на использовании гипотезы компактности. В литературе по распознаванию образов [7] для определения компактности вводится понятие граничного индекса. Пусть *Q* – произвольное множество из *m* объектов. Найдутся такие m_1 объектов, в любой окрестности которых находятся как

точки, принадлежащие данному множеству Q, так и не принадлежащие ему. Эти m_1 объектов называются граничными. Тогда отношение m_1/m называется граничным индексом. Чем он меньше, тем выше компактность множества Q. Считается, что образом может называться такое множество, которое обладает достаточно малым значением граничного индекса.

Простому образу соответствует компактное множество точек в пространстве характеристик, если :

1) значение граничного индекса мало;

2) любые две внутренние точки множества могут быть соединены достаточно плавной линией, проходящей только через точки того же образа и, наоборот, от точки одного образа нельзя перейти к точке другого образа без того, чтобы не пройти точку с неопределенной принадлежностью к образам;

3) почти каждая внутренняя точка образа имеет в достаточно обширной окрестности только точки этого же образа.

Иногда простыми или компактными называются такие образы, которые отделяются друг от друга «не слишком вычурными границами».

Приведенные определения компактности оперируют такими нечеткими понятиями, как «мало», «достаточно плавная линия», «достаточно обширная окрестность», «не слишком вычурная граница». Хотелось бы получить количественную меру компактности, причем такую, значение которой было бы прямо связано с ожидаемой надежностью распознавания.

Одна из мер такого рода предложена в [8] и состоит в вычислении профиля компактности. Пусть для объекта a обучающей выборки все остальные (M-1) объектов упорядочиваются по их расстоянию до a. Анализируются такие порядки, построенные для всех объектов. При движении от самого близкого соседа a до самого далекого его соседа для каждой порядковой позиции 1, 2, …, i,…(M-1) определяется количество объектов m_i, которые не принадлежат тому образу, которому принадлежит объект a. Величины $V_i = m_i/m$ и формирует профиль компактности. Чем компактнее образы, тем для большего числа первых порядковых номеров i величина $V_i = 0$. Если образы сильно перекрывают друг друга, то профиль компактности будет представлять собой ломаную кривую, колеблющуюся около величины V_i =0,5. Если образы удаляются друг от друга, то начальный участок кривой $V_i = f(i)$ все больше приближается к 0, а конечный участок – к величине $V_i = 1$. Профиль компактности дает наглядное качественное представление о ситуации. Переход от графика функции $V_i = f(i)$ к количественной оценке компактности может делаться разными способами, но этот вопрос в [8] не рассматривается.

Нам представляется, что для получения количественной оценки компактности каждого образа в отдельности и качества (информативности) признакового пространства можно использовать FRiS-функцию, описанную выше.

В результате выполнения пунктов 0—8 в алгоритме FRiS-Stolp мы получаем значение величины F, характеризующей защитные и толерантные свойства каждого объекта обучающей выборки. Просуммировав величины F_i всех объектов образа A и разделив сумму на M_A, мы получим среднее значение G_A этих величин у образа A.

$$G_A = \frac{1}{M_A}\sum_{i=1}^{M_A} F_i$$

Аналогичным способом можно получить и величину G_B для образа B. Эти величины обладают следующим свойством. Если расстояния между образами велики и образы представляют собой плотные («компактные») сгустки объектов, то расстояния от объектов до своих столпов будут существенно меньшими по сравнению с расстояниями до столпов образа-конкурента, и значения величин G_A и G_B будут стремиться к 1. Если образы будут приближаться друг к другу, эти величины будут уменьшаться. Если образы начнут пересекаться, то для части объектов ближайшим соседом может оказаться объект

образа-конкурента, и функция сходства таких объектов со своим столпом станет отрицательной, в результате величины G_A и G_B могут уменьшиться и перейти нулевое значение. Если образы будут наложенными друг на друга так, что их объекты будут перемешанными по типу «губка-вода», то значение функции сходства будет отрицательным для большинства объектов и величины G_A и G_B будут стремиться к -1.

При больших расстояниях между образами граница между образами будет представлять собой простейшую поверхность (гиперплоскость). О такой ситуации можно говорить, что образы обладают хорошей компактностью, и можно ожидать высокой надежности распознавания контрольных объектов. Если образы пересекаются настолько, что граничных объектов будет много, а разделяющая поверхность будет очень сложной, и величины G_A и G_B близки к 0, то это означает, что компактность сильно разрушена и результаты распознавания будут низкими. При полном наложении образов друг на друга большинство объектов станут граничными, и граница между образами будет невообразимо сложной. Образы полностью потеряют компактность, результаты распознавания будут носить случайный характер.

Мы видим, что величины G_A и G_B хорошо коррелируют с интуитивными представлениями о компактности образов: чем выше компактность, тем больше эти величины и тем выше ожидаемые результаты распознавания. И наоборот, с уменьшением компактности уменьшаются и они.

Общая оценка компактности K образов в данном признаковом пространстве, а, следовательно, и информативности этого пространства, может быть получена путем арифметического или геометрического усреднения оценок G_j. Если нам нужно найти признаки, наиболее информативные для всех образов в среднем, тогда общий критерий информативности G' должен быть таким:

$$G' = (1/K)\sum_{j=1}^{K} G_j$$

Если мы стремимся к тому, чтобы компактность самого некомпактного образа G_j была максимально возможной, тогда нужно выбрать такую подсистему признаков, при которой достигает максимума следующая величина:

$$G = \sqrt[K]{\prod_{j=1}^{K} G_j}$$

Чем выше плотность объектов внутри образов и чем дальше образы отстоят друг от друга, тем больше величина компактности. Таким же свойством обладает и мера, предложенная Фишером для оценки информативности признаков. Различие состоит в том, что мера Фишера предназначена для образов с нормальным распределением объектов, а мера компактности применима для произвольных распределений. Вполне естественным является использование компактности в качестве **критерия информативности** признакового пространства. Наши эксперименты с этим критерием показали его существенное преимущество по сравнению с широко используемым критерием минимума ошибок при распознавании тестовой выборки [10].

6. Распознавание двух видов лейкемии – ALL и AML

Адекватность эмпирических гипотез, лежащих в основе описанного выше FRiS-подхода, может быть подтверждена или опровергнута результатами решения модельных или хорошо изученных реальных задач. Ниже будет описан опыт применения FRiS для решения задачи распознавания двух типов лейкемии. Эта задача интересна тем, что в литературе представлены результаты ее решения разными группами исследователей. В частности, в работе [11] описаны результаты, которые на момент публикации были лучшими в мире. Они получены с использованием метода Support Vector Machine (SVM),

высокая эффективность которого подтверждена результатами решения большого количества трудных задач. Это дает возможность сравнить наши результаты с лучшими прежними результатами.

Анализируемые данные представлены матрицей векторов экспрессии генов, полученных с помощью биочипов для пациентов с двумя типами лейкемии – AAL и AML [11,12]. Обучающая выборка, полученная на образцах костного мозга, содержит 38 объектов (27 All и 11 AML). Тестовая выборка имеет 34 объекта (20 ALL и 14 AML), которые получены в разных экспериментальных условиях: 24 на препаратах из костного мозга и 10 на препаратах из крови. Исходное количество признаков (генов) N=7129. Нормализованные уровни экспрессии генов измерены по изображениям биочипов.

Результаты решения этой задачи, описанные в работе [11], таковы. Информативное подмножество признаков выбиралось методом RFE (разновидностью алгоритма Deletion [13]), который состоит в поочередном исключении наименее информативных признаков). Решающие правила основаны на методе SVM. В исходном пространстве 7129 признаков правильно распознавалось 29 контрольных объектов из 34 (здесь и далее приводятся результаты, называемые в [11] Success rate). Затем были найдены наилучшие подсистемы, размерность которых кратна степени числа 2: 4096, 2048, …, 4 и 2. По двум лучшим признакам, которые можно выбрать по результатам обучения, правильно распознано 30 объектов, по 4 лучшим признакам – 31, по 128 признакам – 33. В работе указаны также подсистемы из 2, 8 и 16 признаков, которые правильно распознают все 34 контрольных объекта, но по результатам обучения выбрать их было бы не возможно. Для получения описанных результатов машина класса Пентиум работала три часа.

Нами на тех же данных получены следующие результаты [14]. В исходном признаковом N-мерном пространстве без выбора эталонных объектов (все 38 обучающих объектов считались столпами) правильно распознается P=28 из 34 контрольных объектов. Информативное подмножество признаков выбиралось с помощью алгоритма FRiS-GRAD [10]. Этот алгоритм сначала оценивает каждый признак в отдельности, отбирает подмножество из n<<N наиболее информативных признаков (в данном случае n=100) и из них методом полного перебора строит вторичные признаки («гранулы») в виде наилучших пар и троек признаков. Выбор наилучших сочетаний гранул делается итеративной процедурой «Addition – Deletion» [15]. Информативность отдельных признаков и их сочетаний оценивается по критерию FRiS-компактности. Из исходного количества 7129 признаков этим методом было выбрано 39 признаков, из которых программа FRiS-Stolp построила 30 вариантов решающих правил. Первые 27 правил дают результат P= 34 из 34. Десять наиболее информативных правил показаны в таблице 1. В состав каждого правила входит от четырех до шести признаков с весами, которые указаны под косой чертой. Высокие оценки компактности G образов в выбранных подпространствах говорят о большой их информативности и высокой надежности принятых решений.

Таблица 1. Правила принятия решений.

№	Решающие правила	G	P
1	356/1 + 2266/1 + 2358/1 + 2641/5 + 4049/5 + 6280/1	0,73835	34
2	356/1 + 2266/1 + 2358/1 + 2641/4 + 2724/1 + 4049/4	0,73405	34
3	356/1 + 2266/1 + 2641/4 + 3772/1 + 4049/4 + 4261/1	0,73302	34
4	1383/1 + 1833/1 + 2641/4 + 4049/4 + 5441/1 + 6800/1	0,73263	34
5	356/1 + 435/1 + 2641/4 + 4049/4	0,73214	34
6	356/1 + 435/1 + 2641/4 + 2724/1 + 4049/4	0,73204	34
7	1833/1 + 2641/4 + 4049/4 + 4367/1 + 4873/1 + 6800/1	0,73088	34
8	356/1 + 435/1 + 2641/4 + 3560/1 + 4049/4 + 6800/1	0,72919	34
9	356/1 + 2641/4 + 2895/1 + 3506/1 + 4049/4 + 5059/1	0,72814	34
10	356/1 + 2266/1 + 2641/4 + 4049/4 + 4229/1 + 6280/1	0,72699	34

Из таблицы 1 видно, что во всех правилах присутствуют признаки 2641 и 4049. Два этих признака дают результат 33 из 34. Проекция обучающих и контрольных объектов на плоскость двух этих признаков показана на рис. 3. Для одного образа (ALL) потребовался один столп, для другого (AML) – два столпа.

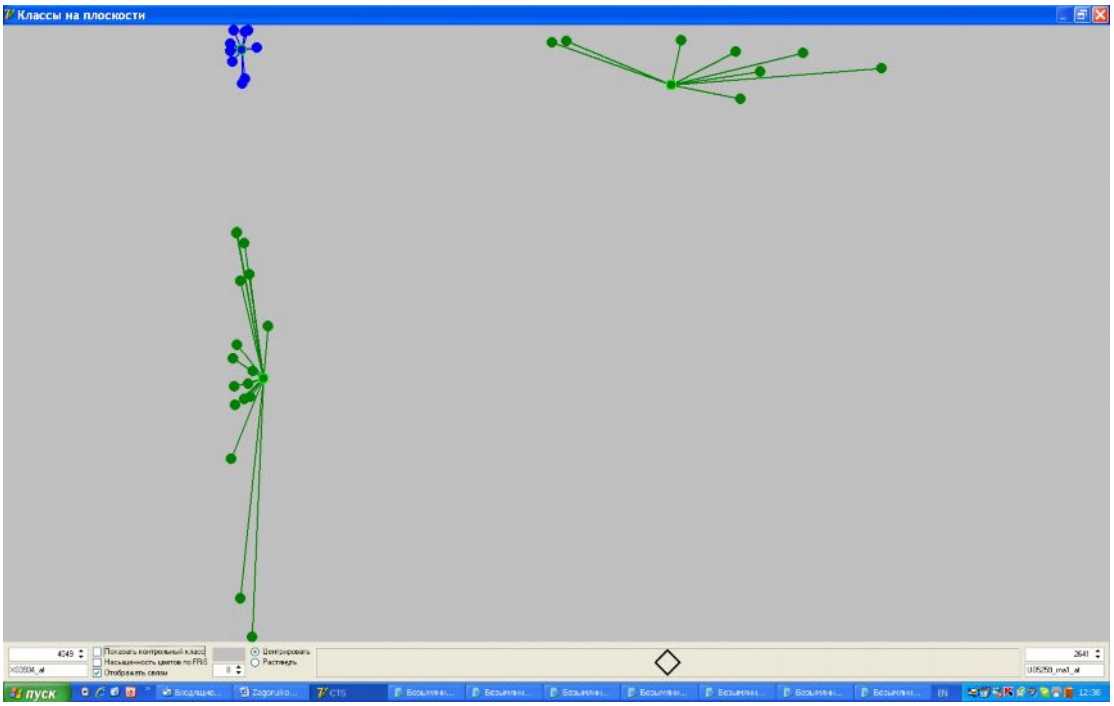

Рис. 3. Объекты обучающей выбороки классов ALL (слева вверху) и AML (справа и внизу) в проекции на признаки 2641 и 4049.

Трудоемкость алгоритма равна $C*(N+n^3/6)*M^3$. где N – исходное количество признаков, $n<<N$ - количество признаков, из которых формируются гранулы, M – количество объектов обучающей выборки. Машинное время, практически, не зависит от исходного количества признаков N и сильно растет с ростом количества обучающих объектов M. В рассматриваемой задаче M было невелико, и описанное выше решение на Пентиуме получено за 50 сек.

Различие между приведенными результатами и результатами работы [11] могут зависеть от двух причин: от метода выбора признаков и типа решающих правил. Для сравнения решающих правил SVM и FRiS был проведен такой эксперимент. В подпространстве двух признаков (генов 803 и 4846), выбранных методом FRE, по правилу SVM было получено 30 правильных ответов, а FRiS-методом -33. По одному гену (4846) результат SVM равен 27, а результат FRiS равен 30. Сравнение двух методов выбора признаков FRE и FRiS-GRAD показывает следующее: Для двух лучших признаков, выбранных методом FRiS-GRAD (2641 и 4049) FRiS-правило дает 33 правильных ответа, а по одному гену (2461) – 32 правильных ответа. Отсюда можно сделать вывод, что, как методы выбора признаков, так и решающие правила, основанные на FRiS- функции, обладают достаточно высокими конкурентными качествами.

7. Прогнозирование значения признака, измеренного в абсолютной шкале

Достаточно успешным оказалось применение FRiS-функции при решении задачи на международном конкурсе Data Mining Cup 2009 [15]. Задача состояла в предсказании значений переменных, измеренных в абсолютной шкале, и заключалась в следующем. Анализировались данные о том, сколько книг того или иного жанра было продано в разных магазинах в течение года. Эти данные представляли собой очень разреженную таблицу (84% клеток таблицы были пустыми), в которой M строками (объектами) являлись магазины ($M = 4812$), а N столбцами (признаками) — жанры книг ($N = 1864$). На пересечении строки и столбца указывалось количество книг данного жанра, проданных в течение года в данном магазине. Признаки имели значения от 0 до 2300. Последние 8 признаков объявлены целевыми. Таблица была разделена по горизонтали на два слоя. В первом (обучающем) слое содержалось $M_o = 2394$ магазина. Для них были указаны значения как описывающих, так и целевых признаков. Во втором слое из $M_k = 2418$ магазинов содержалась информация только об $(N - 8)$ описывающих признаках. Для этих контрольных магазинов требовалось предсказать (угадать), сколько и каких из 8 жанров книг было продано в каждом из них. Это означает, что нужно предсказывать значения целевых признаков, измеренных в абсолютной шкале, в 19344 ячейках матрицы размером 2418×8. Качество решения оценивалось суммой модулей разностей между фактическими и предсказанными значениями в каждой ячейке.

Переход от предсказания в номинальной шкале (распознавание образов) к предсказанию в абсолютной шкале (прогонозирование) потребовал разработки новых схем использования FRiS-функции, что привело к созданию алгоритма FRiS-Pro [16]. Алгоритм использует сходство между профилями строк, что потребовало нормировки строк по их средним значениям. Обучение и распознавание делалось для каждого целевого признака в отдельности. При оценке информативности описывающего признака на данных обучающей подтаблицы для каждой строки в режиме One-Leave-Out (OLO) определяется значение целевого признака, которое принимается равным средневзвешенному значению целевого признака у k ближайших соседей этой строки. При этом используются разные варианты параметров алгоритма — способы нормировки и усреднения, виды метрики, число соседей k и т. д. Разница между предсказанным и истинным значениями суммируется в счетчике штрафа. В итоге процедуры OLO получается оценка информативности данного описывающего признака. Такой же способ оценки применяется и при выборе гранул признаков и формировании решающих правил.

Значение FRiS-функции используется для оценки весов k ближайших соседей, участвующих во взвешенном усреднении целевого признака. Расстояние от контрольного объекта z (строки) до каждого из k ближайших соседей (объектов «своего» класса) играли роль расстояний *R1*, а за расстояние *R2* до столпа класса-конкурента принималось среднее расстояние от z до следующих по порядку k ближайших соседей. По этим расстояниям вычислялось значение FRiS-функции для каждого из k ближайших соседей. Предсказываемое значение целевого признака у объекта z получалось в результате взвешенного усреднения значений этого признака у k «своих» ближайших соседей. При усреднении вес каждого из этих соседей был равен значению его FRiS-функции.

После обучения фиксировалось наилучшее сочетание значений параметров алгоритма для каждого целевого признака в отдельности, и делалось предсказание целевых признаков контрольной части таблицы.

В конкурсе изъявили желание участвовать 618 команд из 164 организаций 42 стран. 231 команда решила эту задачу и прислала свои результаты. 49 команд преодолели порог приемлемых результатов, установленный организаторами. Результаты первых 10 и некоторых других команд приведены в таблице 2.

Таблица 2. Результаты решения задачи прогнозирования.

1	Uni Karlsruhe (TH)_ II	17260
2	TU Dortmund	17912
3	TU Dresden	18163
4	**Novosibirsk State University**	**18353**
5	Uni Karlsruhe (TH)_ I	18763
6	FH Brandenburg_I	19814
7	FH Brandenburg_II	20140
8	Hochschule Anhalt	20767
9	Uni Hamburg_	21064
10	KTH Royal Institute of Technology	21195
11	RWTH Aachen_I	21780
14	Budapest University of Technology	23277
15	Isfahan University of Technology	23488
16	TU Graz	23626
18	Uni Weimar_I	23796
19	Zhejiang University of Sc. and Tech	23952
20	University Laval	24884
24	University of Southampton	25694
25	Telkom Institute of Technology	25829
26	University of Central Florida	26254
32	Indian Institute of Technology	28517
34	Anna University Coimbatore	28670
38	Technical University of Kosice	32841
39	University of Edinburgh	45096
48	Warsaw School of Economics	77551
49	FH Hannover	1938612

Как видно из этой таблицы, среднее количество ошибок на одну предсказываемую ячейку у разных команд колебалось от 0.89 до 100.22. Наша команда Новосибирского Университета сделала 0.95 ошибки на ячейку и заняла 4 место. Полученные результаты подтверждают возможность использования FRiS-функции в алгоритмах решения задач прогнозирования количественных переменных.

Кроме этой задачи описанными методами были успешно решены и другие задачи распознавания из области медицины (раковые заболевания по масспектрам белков [17]), физики (классификация мелкодисперсных веществ по рентгеновским спектрам [18]) и т.д. Общее свойство этих задач состояло в том, что отсутствовала информация о характере распределений образов и о зависимостях признаков, а количество признаков N на порядки превышало количество обучающих объектов M.

Заключение

Использование относительной меры сходства, учитывающей конкурентную обстановку, позволяет строить эффективные алгоритмы решения всех основных задач Data Mining. Функция конкурентного сходства (FRiS) позволяет получать количественную оценку компактности образов и информативности признакового пространства и строить легко интерпретируемые решающие правила. Метод применим к задачам с произвольным количеством образов, любому характеру их распределений и обусловленности обучающей выборки (соотношению между M и N). Трудоемкость метода позволяет использовать его для решения достаточно сложных реальных задач. Качество решения прикладных задач не уступает качеству, получаемому другими методами.

Продолжение исследований FRiS – подхода предусматривает его распространение на другие типы задач анализа данных (заполнение пробелов, поиск ассоциаций и т.д), на исследование других типов функции конкурентного сходства, на решение проблемы цензурирования обучающей выборки и пр.

Благодарности

Работа была выполнена при поддержке Российского фонда фундаментальных исследований, грант 08-01-00040 и Международного фонда «Научный потенциал».

Литература:

1. Воронин Ю.А. Начала теории сходства. Изд. ВЦ СО РАН, Новосибирск, 1989.
2. Ю.А. Шрейдер. Равенство, сходство, порядок. «Наука», М. 1971 г.
3. E. Fix and J. Hodges. Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties. Technical report, USAF School of Aviation Med., Randolph Field, TX, Rep. 21-49-004, 1951.
4. Kira K., Rendell L. The Feature Selection Problem: Traditional Methods and a New Algorithm // Proc. 10th Nat'l Conf. Artificial Intelligence (AAAI-92). – 1992. P. 129-134.
5. 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.1pp.1-6.
6. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Изд. ИМ СО РАН, Новосибирск, 1999. 270 с.
7. Braverman E.M. Experiences on training the machine to recognition of visual patterns // Automatics and Telemechanics. 1962, Vol. 23, №3, pp. 349-365.
8. Воронцов К. В., Колосков А. О. Профили компактности и выделение опорных объектов в метрических алгоритмах классификации // Искусственный Интеллект. 2006. С. 30–33.
9. Nikolay Zagoruiko. Measure of Similarity and Compactness in Competitive Space // Advances in Intelligent Data Analysis VIII. Springer-Verlag Berlin Heidelberg 2009. Printed in Germany
10. N.G. Zagoruiko, O.A. Kutnenko and A.A. Ptitsin. Algorithm GRAD for Selection of Informative Genetic Feature // Proc. Int. Conf. on Computational Molecular Biology, Moscow. 2005. pp.8-9.
11. Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002, 46 (1-3): pp. 389-422.
12. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. // Science, Vol 286, Oct 1999.
URL: http://www.genome.wi.mit.edu/MPR/data_set_ALL_AML.html
13. Merill T., Green O.M. On the effectiveness jf receptions in recognition systems // IEEE Trans. Inform. Theory. 1963. V. IT-9, p. 11-17.
14. Борисова И.А., Дюбанов В.В., Загоруйко Н.Г., Кутненко О.А. Сходство и компактность. Труды 14-й Всероссийской Конференции «Математические методы распознавания образов». Г. Суздаль, 2009 г.
15. http://www.prudsys.de/Service/Downloads/bin/ DMC2009_Ergebnisliste.pdf
16. Применение FRiS-функции для решения задачи прогнозирования спроса.(алгоритм FRiS-Pro). Труды Всероссийской с международным участием Конференции «Знания-Онтологии-Теории» (ЗОНТ-09). Новосибирск, 2009.
17. N.G. Zagoruiko and O.A. Kutnenko. Recognition Methods Based on the AdDel Algorithm // Int. Journal "Pattern Recognition and Image Analysis", 2004, vol. 14, № 2, pp.198-204.

18. 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, №1, 2009.сс. 92-101.