Публикация: Труды Всероссийской Конференции «Знания-Онтологии-Теории» (ЗОНТ-09), Новосибирск, 2009, Том I, с. 93-102.

Знания-Онтологии-Теории (ЗОНТ-09)

Меры сходства, компактности, информативности и однородности обучающей выборки

Загоруйко Н. Г.[1], Борисова И. А.[1], Дюбанов В. В.[2], Кутненко О. А.[1]
[1]*Институт математики им. С.Л. Соболева СО РАН,* [2]*Новосибирский Государственный Университет*
zag@math.nsc.ru

Аннотация. *В статье описывается, как с помощью функции конкурентного сходства (FRiS-функции) можно оценивать сходство между объектами и образами, получать количественные меры компактности образов, информативности признакового пространства и однородности обучающей выборки. Представлен опыт использования предлагаемых мер для решения задач распознавания и прогнозирования количественной переменной.*

Ключевые слова: *функция конкурентного сходства, распознавание, однородность, компактность, прогнозирование, информативность, цензурирование.*

1 Мера конкурентного сходства

Сходство $S(a,b)$ двух объектов a и b обычно оценивается величиной, которая зависит от расстояния $R(a,b)$ между этими объектами и обладает свойствами симметричности, рефлексивности и неравенства треугольника. 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 $\overline{S}$ не обладает свойством симметричности: $\overline{S}(a,b) \neq \overline{S}(b,a)$. Не всегда выполняется для этой меры и неравенство треугольника: сумма сходств $\overline{S}(a,b)+\overline{S}(a,c)$ может быть меньше сходства $\overline{S}(b,c)$. 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 известном 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

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

где $R_{\max}$ и $R_{\min}$ — максимальное и минимальное расстояния между всеми парами объектов. Нормализация разности расстояний по величине $(R_{\max} - R_{\min})$ представляется неудачной. 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.

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 — минимальное 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 («**F**unction of **Ri**val **S**imilarity») удовлетворяет всем этим требованиям [2]:

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

Абсолютная мера сходства $S = 1 - R$ (здесь R — нормированное от 0 до 1 расстояние) дает трудно интерпретируемый ответ на вопрос: «Насколько объект z похож на эталон 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 эти вопросы и, кроме того, на такой вопрос: «Насколько величина сходства объекта z с эталоном образа A больше сходства z с самым сильным конкурентом – эталоном образа B?»

2 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.

Для произвольного объекта a_i, $i \in \{1,\ldots,M_A\}$, образа A вычисляются две характеристики C_i^1 и C_i^2 эффективности этого объекта в роли единственного столпа данного образа. В качестве столпов конкурирующего образа B используются все его M_B объектов.

1. Для произвольного объекта a_j, $j \neq i$, образа A находим расстояние $R(a_j,a_i)$ до столпа a_i и расстояние $R(a_j,b)$ до ближайшего к нему объекта b образа B. По этим расстояниям вычисляем значение функции сходства:

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

Чем больше эта величина, тем лучше объект a_i защищает объект a_j от включения его в состав образа B. Если величина $F_{a_i/b}(a_j)$ больше порога F^*, например, больше 0, то будем считать, что объект aj входит в состав кластера, образуемого объектом ai, а величину $F_{a_i/b}(a_j)$ добавим к счетчику C_i^1.

2. Повторив шаг 1 для всех объектов образа A, получим в счетчике C_i^1 сумму оценок сходства тех m_i объектов образа A, которые вошли в состав кластера, образованного объектом a_i. Разделив эту сумму на m_i, получим оценку F_i^1 «обороноспособности» объекта a_i:

$$F_i^1 = \frac{C_i^1}{m_i}.$$

3. 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. Для произвольного объекта b_q вычислим величину $F_{a_i/b_s}(b_q) = \dfrac{R(b_q,a_i) - R(b_q,b_s)}{R(b_q,a_i) + R(b_q,b_s)}$ его сходства 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.

5. Повторив шаг 4 для всех объектов образа B, 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:

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

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 1-5, получим такие оценки для всех M_A объектов образа A. В качестве столпа первого кластера образа A выбираем тот объект a_i, которому соответствует наибольшая величина F_i.

7. Если не все M_A объектов вошли в этот кластер, то для остальных объектов повторяем шаги 1–6 до тех пор, пока все объекты обучающей выборки образа A не окажутся включенными в свои кластеры. В итоге образ A будет представлен k_A столпами, k_A>=1.

8. Затем выполним шаги 1–7 для объектов b_s, $s = 1,\ldots,M_B$, образа B, в результате чего получим k_B столпов образа B, k_B>=1.

9. Для проверки устойчивости полученного решения повторим шаги 1–8 с той разницей, что в качестве столпов конкурирующих образов будем использовать не все их объекты, а лишь столпы, выбранные на предыдущем этапе. Опыт показывает, что, как правило, одной такой проверки оказывается достаточно.

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).

Алгоритм FRiS-Stolp обладает следующими свойствами. 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, а значение функции сходства объекта с выбранным образом позволяет судить о достоверности принятого решения.

3 Оценка компактности и информативности

Практически все алгоритмы распознавания основаны на использовании гипотезы компактности [3]. В соответствие с этой гипотезой простому образу соответствует компактное

множество точек в пространстве характеристик, если 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 «почти каждая точка», «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 [4] и состоит в вычислении профиля компактности. Пусть для всех M объектов a обучающей выборки все остальные $(M-1)$ объектов упорядочиваются по их расстоянию до a. При движении вдоль этих упорядоченных списков от первой позиции $j=1$ до последней $j=M-1$ в каждой порядковой j-той позиции определяется количество объектов m_j, которые не принадлежат тому образу, которому принадлежит объект a. Усредненные по всей выборке 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 j профиля выполнено равенство $V_j = 0$. Переход от профиля к количественной оценке компактности может делаться разными способами. В работе [4] описывается связь между профилем компактности и функционалом полного скользящего контроля, который является естественной количественной оценкой компактности для метода kNN.

Для получения количественной оценки компактности мы предлагаем использовать описанную выше FRiS-функцию. Будем оценивать компактность образа A в задаче распознавания K образов. При выборе столпов образа A для всех его объектов были получены оценки $\overline{F}_i$, которые тем больше, чем больше сходство с ними остальных объектов своего образа и чем меньше сходство с ними объектов чужих образов. 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 $\overline{F}_i$ будут стремиться к 1. Если образы будут приближаться друг к другу, величины $\overline{F}_i$ будут уменьшаться. Если образы начнут пересекаться, то компактность начнет разрушаться и некоторые объекты окажутся в окружении чужих объектов. Они получат отрицательное значение величины $\overline{F}_i$. Если образы будут наложенными друг на друга так, что их объекты будут перемешанными по типу «губка-вода», то это означает, что компактность разрушена полностью и значения $\overline{F}_i$ почти у всех объектов будут отрицательными.

Мы видим, что величины $\overline{F}_i$ хорошо коррелируют с интуитивными представлениями о компактности (разделимости) образов: чем выше компактность, тем больше величины $\overline{F}_i$, и тем выше ожидаемые результаты распознавания. И, наоборот. На этом основании компактностью G_A образа A будем считать величину, зависящую от среднего значения оценок $\overline{F}_i$:

$$G_A = \frac{1}{M_A}\sum_{i=1}^{M_A}\overline{F}_i .$$

Общая оценка G компактности K образов в данном признаковом пространстве может быть получена путем арифметического или геометрического усреднения. Для минимизации ошибок всех образов в среднем следует использовать арифметическое усреднение:

$$G' = \frac{1}{K}\sum_{j=1}^{K} G_j .$$

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

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

Наши эксперименты показывают, что критерий G обычно дает лучший результат по сравнению с критерием G'. Описанная мера компактности тем больше, чем выше плотность объектов внутри образов и чем дальше образы отстоят друг от друга. 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. Она используется в этом качестве в алгоритме FRiS-GRAD [5]. Наши эксперименты с этим критерием показали его существенное преимущество по сравнению с широко используемым критерием минимума ошибок при распознавании тестовой выборки методами Cross Validation или One Leave Out[2].

4 Оценка однородности и цензурирование выборки

Найденные в процессе выбора столпов оценки $\overline{F_i}$ каждого объекта позволяют анализировать однородность обучающей выборки. В качестве меры однородности выборки объектов образа будем использовать дисперсию d значений $\overline{F_i}$ их сходства со своими столпами. Оценка $\overline{F_i}$ у объекта, находящегося в центре локального сгустка своих объектов, будет больше, чем у периферийных объектов. Для объектов, оказавшихся в окружении чужих объектов, величина $\overline{F_i}$ может иметь отрицательное значение. Такие объекты (выбросы — «outliers») будут приводить к увеличению числа столпов и ухудшению последующего качество распознавания. По этой причине их целесообразно исключить из дальнейшего рассмотрения («цензурировать»).

Процесс цензурирования состоит из последовательного исключения объектов и пересчета значений $\overline{F_i}$ для оставшихся объектов. Сначала исключается объект, обладающий наименьшим значением величины $\overline{F_i}$. После пересчета можно увидеть, что дисперсия d уменьшилась. Одновременно выявляется другой объект с минимальным значением $\overline{F_i}$, который является кандидатом на очередное исключение.

Если этот процесс не останавливать, то минимальная дисперсия $d = 0$ будет достигнута, когда в выборке останутся только объекты-столпы. Цензурирование должно остановиться на шаге, при котором достигает максимума критерий Q_d, отражающий два противоречивых желания: добиться минимального значения дисперсии d и максимального сохранения количества объектов обучающей выборки:

$$Q_d = f\left(\frac{M_c}{d * M}\right).$$

Здесь Mc/M – доля объектов обучаю.щей выборки, оставшихся в составе выборки после очередного шага цензурирования. В настоящее время исследуются разные варианты критерия остановки процесса цензурирования. В работающих программах распознавания пока используется простой вариант цензурирования: столпы строятся на полной выборке, выделяются «единичные» столпы, которые защищают только самих себя, и эти столпы игнорируются (исключаются из списка столпов) в процессе распознавания контрольных объектов.

5 Примеры решения реальных задач

Описанные выше меры сходства, компактности, информативности и однородности обучающей выборки позволяют унифицировать подходы к решению разных задач распознавания образов. На их основе разработаны алгоритмы построения решающих правил (FRiS-Stolp) и таксономии (FRiS-Class [5]), алгоритм комбинированного типа DX для одновременного построения решающих правил и выбора признаков (FRiS-GRAD [6]) и алгоритм универсальной классификации, решающий задачу таксономии-распознавания при любом соотношении классифицированных и не классифицированных объектов в анализируемой выборке [7]. Ниже

показаны примеры применения некоторых из этих алгоритмов при решении реальных задач. Их общая особенность состоит в том, что количество признаков на порядки превышает количество объектов.

5.1 Диагностика рака простаты по масс-спектрам белков

Анализируются данные о масс-спектре белковых форм, полученные с помощью спектрометра типа SELDI-MS-TOF [8]. Количество признаков (спектральных полос) — 15153. Представлены четыре класса пациентов с разным уровнем индекса PSA, характеризующего степень развития рака простаты: 63 здоровых пациента класса B имеют $\mathrm{PSA} < 1$ ng/mL, 26 пациентов класса C имеют $\mathrm{PSA} = 4 \div 10$ ng/mL, 43 пациента класса D меют $\mathrm{PSA} > 10$ ng/mL и 190 пациентов класса A имеют $\mathrm{PSA} > 4$ ng/mL. Малое количество пациентов не позволяет разделить выборку на обучающую и контрольную. По этой причине воспользуемся тем обстоятельством, что целевая характеристика (PSA), указывающая на принадлежность пациентов к тому или иному классу, позволяет установить между классами отношение частично-линейного порядка.

Если упорядочить классы пациентов по степени проявления 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. Следовательно, 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. В результате можно обучаться на двух классах, а на контроль предъявлять объекты третьего класса. 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 все три класса больных пациентов — классы A, C и D. С помощью алгоритма FRiS-GRAD 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 следующее. Если правила построены для различения класса B (здоровые) от класса C, класса D или их смеси, а на контроль подавать классы больных, не участвовавших в обучении, то класс B хорошо отличается от всех классов больных пациентов (эксперименты 1–5). После обучения на классах C и D выяснилось, что в большом классе A пациентов с признаками принадлежности к классу C («слабо больные») существенно больше, чем пациентов класса D («сильно больные») (эксперимент 6). Если ставить вопрос о том, на какой из классов больных — C или D — больше похожи здоровые люди (эксперимент 7), то ответ будет в пользу класса C («слабо больные»), что вполне естественно.

Таблица 1. Результаты экспериментов

№	Обучение	Контроль	B	C	D
1	B против D	A_{190}	3		187
2	B против D	C_{26}	0		26
3	B против C	A_{190}	1	189	
4	B против C	D_{43}	3	40	
5	B против $(C + D)$	A_{190}	19	137	34
6	C против D	A_{190}		168	22
7	C против D	B_{63}		49	14

Из сказанного выше можно сделать вывод о том, что выбранные 24 спектральных полосы (из 15153) несут информацию, достаточную для получения вполне правдоподобных результатов диагноза пациентов.

5.2 Распознавание двух видов лейкемии — ALL и AML

Задача распознавания двух типов лейкемии интересна тем, что в литературе представлены результаты ее решения разными группами исследователей. В частности, в работе [9] 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 — ALL и AML [10]. 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 [9], таковы. Информативное подмножество признаков выбиралось методом RFE (разновидностью алгоритма Deletion [11], 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 [9] 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 невозможно.

Нами на тех же данных получены следующие результаты. 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 [6]. 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». 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 34 из 34. Десять наиболее информативных правил показаны в таблице 2. В состав каждого правила входит от четырех до шести признаков с весами, которые указаны под косой чертой. В этих правилах задействовано 18 разных признаков. Высокие оценки компактности G образов в выбранных подпространствах говорят о большой их информативности и высокой надежности принятых решений.

Таблица 2. 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 2 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. 1. Для одного образа (ALL) потребовался один столп, для другого (AML) – два столпа.

Трудоемкость алгоритма имеет порядок $O(N + n^3/6)M^3$, 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 растет с ростом количества обучающих объектов со скоростью M^3. В рассматриваемой задаче M было невелико, и описанное выше решение на четырехядерном Пентиуме получено за 30 сек.

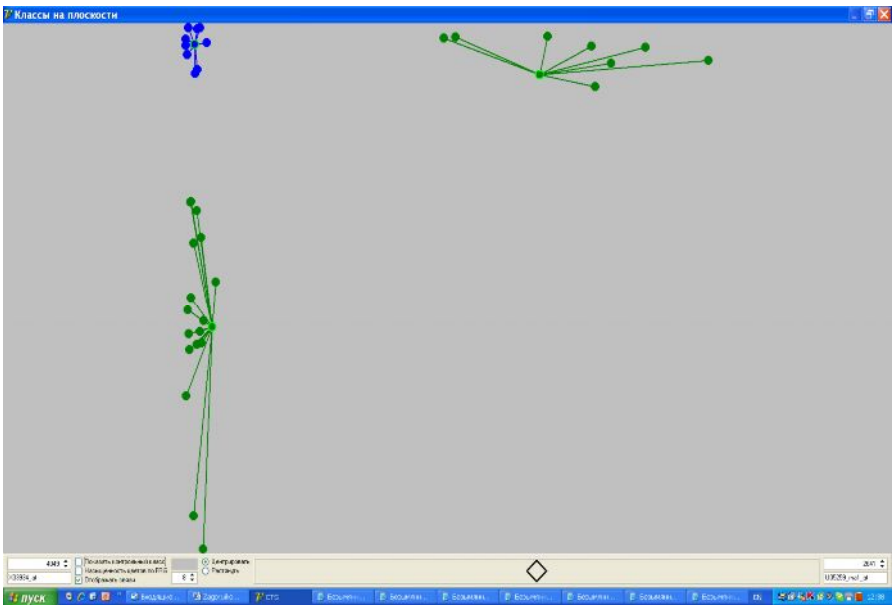

Рис. 1. Объекты обучающей и контрольной выборок классов ALL (слева вверху) и AML (справа и внизу) в проекции на признаки 2641 и 4049.

5.3. Прогнозирование спроса

Достаточно успешным оказалось применение FRiS-функции при решении задачи на международном конкурсе Data Mining Cup 2009 [12]. 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 строк и столбцов указывалось количество книг данного жанра, проданных в течение года в том или ином магазине. Признаки имели значения от 0 до 2300. Последние 8 признаков являлись целевыми. Таблица была разделена по горизонтали на два слоя. В первом (обучающем) слое содержалось $M_o = 2394$ магазина. Для этих магазинов 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 [13]. 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 с помощью метода One-Leave-Out (OLO) определяется значение целевого признака, которое принимается равным среднему 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 r_1, а за расстояние r_2 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 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 3.

Таблица 3. 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.

6 Заключение

Рассмотрение относительной меры сходства, учитывающей конкурентную обстановку, позволяет строить эффективные алгоритмы решения всех основных задач Data Mining, в том числе задач распознавания образов. Функция конкурентного сходства дает возможность вычислять количественную оценку компактности образов и информативности признакового пространства, и строить легко интерпретируемые классификации и решающие правила. Применение FRiS-функции создает предпосылки для решения проблемы цензурирования обучающей выборки. 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. Качество получаемых решений находится на уровне качества лучших опубликованных результатов.

7 Благодарности

Данная работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, проект № 08-01-00040, Международного фонда «Научный потенциал» и гранта АВЦП Рособразования, проект № 2.1.1/3235.

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. —P. 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. — P. 1-6.

[3] *Браверманн Э. М.* Опыт по обучению машины распознаванию зрительных образов // Автоматика и телемеханика —1962. — Т. 23, №3. — с. 349-365.

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

[5] *Борисова И.А.* Кластеризация с использованием функции конкурентного сходства // Научный вестник НГТУ. —2007. — №3(28). — С. 3-12.

[6] *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. — P. 335-338.

[7] *Борисова И.А., Загоруйко Н.Г.* Алгоритм FRiS-TDR для решения обобщенной задачи таксономии и распознавания // (данный сборник).

[8] *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.

[9] *Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik*. Gene Selection for Cancer Classification using Support Vector Machines // Machine Learning. —2002. — V. 46 (1-3). — P. 389-422.

[10] http://www.genome.wi.mit.edu/MPR/data_set_ALL_AML.html

[11] *Merill T., Green O.M.* On the effectiveness jf receptions in recognition systems // IEEE Trans. Inform. Theory. — 1963. —V. IT-9. — P. 11-17.

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

[13] Дюбанов В.В. Использование FRiS-функции в алгоритмах предсказания количественных переменных (Алгоритм FRiS-Pro) . (Данный сборник)