УДК 519.95

КОНКУРЕНТНОЕ СХОДСТВО КАК УНИВЕРСАЛЬНЫЙ БАЗОВЫЙ ИНСТРУМЕНТ КОГНИТИВНОГО АНАЛИЗА ДАННЫХ

Н.Г. Загоруйко, И.А. Борисова[1], О.А. Кутненко[2], В.В. Дюбанов[3], Д.А. Леванов[4]

Институт Математики им. С.Л. Соболева СО РАН,
Конструкторско-технологический институт вычислительной техники СО РАН,
Новосибирский Государственный Университет, Новосибирск, Россия
[1]*biamia@mail.ru,* [2]*olga@math.nsc.ru*

Институт Математики им. С.Л. Соболева СО РАН, Новосибирск, Россия
[3]*vladimir.dyubanov@gmail.com,* [4]*levanovd@gmail.com*

Аннотация

При решении задач анализа данных (классификации, таксономии, выбора признаков, прогнозирования) человек применяет некий универсальный психофизиологический механизм познания, ключевую роль в котором, по нашему мнению, играют способность оценивать меру сходства между объектами и стремление к максимальной компактности и простоте описания мира в терминах этой меры сходства. Николаем Григорьевичем Загоруйко была предложена модель для оценивания сходства объекта с образом, основанная на учете конкурентной ситуации. В статье определяется функция конкурентного сходства (FRiS-функция) и описываются возможности её использования для оценки компактности и разделимости образов. Эти оценки легли в основу алгоритмов для решения задач построения решающего правила (алгоритм FRiS-Stolp), выбора информативных признаков (алгоритм FRiS-GRAD) и цензурирования (алгоритм FRiS-Censor). Основные идеи и свойства этих алгоритмов, а также результаты их применения к модельным и реальным задачам представлены в данной статье.

Ключевые слова: *когнитивный анализ данных, распознавание, выбор признаков, цензурирование, функция конкурентного сходства.*

Введение

Когнитивный анализ данных направлен на изучение методов, с помощью которых человек извлекает закономерности из информации об окружающем мире и затем использует их для получения новых знаний, для более эффективного принятия решений. Эти исследования служат основой для построения компьютерных систем анализа данных, которые, имитируя простейшие когнитивные способности человека, позволяют решать задачи больших объемов, извлекать закономерности тогда, когда человеческих ресурсов на это не хватает. Ярким примером такой задачи могут служить результаты микрочипирования (microarrays) – технологии измерения активности большого количества генов одновременно, которая активно развивается, в том числе, в связи с задачей установления связей между различными заболеваниями и особенностями функционирования генома пациента. Результатом таких исследований становятся таблицы, содержащие информацию о сотнях пациентов и десятках тысяч генов, обработать которые вручную не представляется возможным.

Анализ данных, как направление кибернетики, начал активно развиваться с середины прошлого века, а в связи с появлением компьютеров и непрерывным ростом их мощностей стал неотъемлемой частью исследований в самых разных областях медицины, биологии, геологии, социологии, экономики и пр. У истоков этого направления наряду с такими учёными как С.А. Айвазян [1], Э.М. Браверманн [2], В.Н. Вапник [3], Ю.И. Журавлев [4],

А.Г. Ивахненко [5], М.И. Шлезингер [6] стоял и Николай Григорьевич Загоруйко [7-9], заслуги которого, как учёного и популяризатора анализа данных, сложно переоценить.

Благодаря усилиям большого количества исследователей к настоящему времени в области анализа данных разработано огромное количество разных методов решения одних и тех же задач. Создаются всё новые и новые методы, ориентированные отдельно на большие и малые выборки, на тот или иной закон распределения, на два образа или на большее их число, на линейно или нелинейно разделимые образы и т. д. Подобное многообразие можно объяснить различием моделей, в рамках которых каждый исследователь формулирует и решает ту или иную задачу анализа данных. Причём высокая специфичность модели, несмотря на точность и математическую обоснованность решений, получаемых в ней, может приводить к тому, что для большинства реальных прикладных задач она не будет работать.

Идея, которую Н.Г. Загоруйко продвигал в последние годы работы, заключалась в том, что модель для решения различных задач анализа данных должна быть универсальной, отражать человеческую способность обработки информации и благодаря этому успешно способствовать решению задачи, которые может ставить человек. Он исходил из гипотезы о том, что когнитивные способности человека основаны на двух ключевых принципах: на использовании специфической меры сходства между объектами и на стремлении к максимальной компактности и простоте описания мира в терминах этой меры сходства. В качестве модели человеческого способа оценки сходства между объектами он предложил использовать функцию конкурентного сходства (FRiS-функцию, от *Function of Rival Similarity*) [10]. Использование FRiS-функции позволяет найти количественную оценку компактности классов. Эти два элемента – FRiS-сходство и FRiS-компактность лежат в основе алгоритмов для решения различных типов задач анализа данных. Примеры их успешного использования для решения прикладных задач самой разной природы подтверждают работоспособность модели конкурентного сходства.

1 Функция конкурентного сходства. Алгоритм FRiS-Stolp

При измерении таких характеристик объектов, как вес, длина, сопротивление и т.п., обычно используются эталонные объекты. Результат измерения определяется свойствами только самого объекта и измеряющего эталона и *не зависит от свойств других объектов*. По этой причине он имеет характер абсолютной величины. Но объекты описываются и такими характеристиками, как «похож–не похож», «близок–далёк», «добрый–злой» и т.д. Эталонов для подобных понятий не существует, два объекта с несовпадающими свойствами могут считаться «сходными» или «не сходными», «близкими» или «далёкими» *в зависимости от свойств других объектов*. Так, на рисунке 1 расстояние между объектами *a* и *z* остается неизменным, но ответы на вопрос «достаточно ли они близки друг к другу, чтобы можно было объединить их в один класс?» для случаев 1, 2, и 3 будут разными.

Рисунок 1 – Иллюстрация относительности сходства объектов *a* и *z*

Чтобы ответить на этот вопрос, нужно знать ответ на вопрос «по сравнению с чем?». Хорошо известная бытовая фраза «всё познается в сравнении» на самом деле отражает фундаментальный закон познания. Адекватная мера сходства должна определять величину сходства, зависящую от особенностей конкурентного окружения объекта *z*. 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*. Следовательно, сходство в распознавании образов является категорией не абсолютной, а относительной.

Все статистические алгоритмы распознавания учитывают конкуренцию между классами. Например, в методе kNN (k-Nearest Neighbors[1]) новый объект *z* распознаётся как объект образа *A*, если среднее расстояние до *k* ближайших объектов этого образа не только мало, но и меньше, чем среднее расстояние до *k* ближайших объектов конкурирующего образа *B*. Оценка сходства в этом алгоритме делается в шкале порядка.

Более сложная мера сходства используется в алгоритме RELIEF [11]. Чтобы определить сходство объекта *z* с ближайшим объектом *a* из образа *A* в конкуренции с ближайшим объектом *b* из образа *B* используется величина, которая учитывает нормированную разницу в расстояниях $r(z, a)$ и $r(z, b)$:

$$W(z,a \mid b) = \frac{r(z,b) - r(z,a)}{r_{\max} - r_{\min}}.$$

Здесь r_{min} и r_{max} – минимальное и максимальное расстояния между всеми парами объектов анализируемого множества.

В работе [12] при оценке «Ширины Силуэта» (Silhouette Width) измеряется среднее расстояние $R(z, A)$ от объекта *z* до всех объектов образа *A*, к которому относят *z*, и расстояние $r(z, b)$ от *z* до ближайшего к нему объекта *b*, не принадлежащему образу *A*. Мера различия (несходства) объекта *z* и объектов кластера *A* принимается равной

$$S(z,A \mid b) = \frac{R(z,A) - r(z,b)}{\max\{R(z,A), r(z,b)\}}.$$

Для вычисления конкурентного сходства объекта *z* с объектом *a* в конкуренции с ближайшим объектом *b* предлагается использовать следующую величину:

$$F(z,a \mid b) = \frac{r(z,b) - r(z,a)}{r(z,b) + r(z,a)}.$$

По мере передвижения объекта *z* от объекта *a* к объекту *b* можно говорить вначале о большом сходстве объекта *z* с объектом *a*, об умеренном их сходстве, затем о наступлении одинакового сходства, равного 0, как с объектом *a*, так и с *b*. При дальнейшем продвижении *z* к *b* возникает умеренное, а затем и большое отличие *z* от *a*. Совпадение объекта *z* с объектом *b* означает максимальное отличие *z* от *a*, что соответствует сходству *z* с *a*, равному -1.

Сходство *F* между объектами не зависит от положения начала координат, поворота координатных осей и одновременного умножения их значений на одну и ту же величину. 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 учитываются разные характеристики. Меняя веса характеристик, можно подчеркнуть сходство или различие между заданными объектами, что обычно и делается при выборе информативных признаков и построении решающих правил в распознавании образов.

[1] *k Nearest Neighbor* (*k Ближайших Соседей)* — один из самых простых алгоритмов классификации, используемый также в задачах регрессии. *Прим. ред.*

Конкурентное сходство объектов с образами будем определять по тому же принципу, что и конкурентное сходство объектов с объектами:

$$F(z, A \mid B) = \frac{r(z,B) - r(z,A)}{r(z,B) + r(z,A)}.$$

При этом расстояние от объекта z до образов A и B может вычисляться по-разному. В качестве него может использоваться и расстояние $r(z, a)$ до ближайшего объекта a образа A, и среднее расстояние до всех объектов образа, и среднее расстояние до k ближайших объектов образа, и расстояние до центра тяжести образа и т.п. В дальнейшем, в качестве расстояния от объекта до образа по умолчанию будет использоваться расстояние до ближайшего объекта этого образа.

Сходство в шкале порядка, используемое в методе kNN, отвечает на вопрос «на объекты какого образа объект z похож больше всего?». Конкурентное сходство, измеряемое с помощью FRiS-функции, отвечает на этот вопрос и, кроме того, на вопрос «какова абсолютная величина сходства z с образом A в конкуренции с образом B?». Оказалось, что дополнительная информация, которую даёт абсолютная шкала по сравнению со шкалой порядка, позволяет существенно улучшить методы анализа данных.

В таблице 1 приводится сравнение различных критериев для выбора информативных признаков. FRiS-критерий сравнивался с такими критериями выбора признаков как, Silhouette Width (SW), ошибка скользящего контроля по правилу ближайшего соседа (NN), критерий Фишера (Fisher) и RELIEF. Сравнение осуществлялось на двух сериях выборок. В первом случае образы были представлены нормальными распределениями, искусственно «испорченными» добавлением зашумленных и случайных признаков (V1). Во втором случае образы изначально имели сложную полимодальную структуру, которая маскировалась добавлением зашумлённых и случайных признаков (V2). Решалась задача выбора информативной подсистемы признаков. Для оценки эффективности того или иного критерия использовалась надёжность распознавания тестовой выборки в выбранном с помощью критерия признаковом пространстве.

Таблица 1 – Сравнение различных критериев для выбора информативных признаков

	V1			V2		
\N	50	100	150	50	100	150
FRiS	**0.930308**	**0.938042**	**0.939497**	**0.818071**	**0.856152**	**0.866458**
SW	0.924218	0.937614	0.943244	0.655248	0.690778	0.702515
NN	0.895255	0.894467	0.906625	0.735635	0.772785	0.80093
Fisher	0.935972	0.943948	0.942755	0.598515	0.60059	0.605303
RELIEF	0.899302	0.932148	0.930718	0.710267	0.754834	0.791442

Анализ представленных результатов позволяет заключить, что если в простых случаях унимодальных образов с небольшой зоной пересечения все критерии работают примерно одинаково, то в случае сложных полимодальных структур FRiS-критерий значительно опережает другие критерии.

Одним из способов проявить особенности данных в задаче распознавания является переход к их сжатому описанию с помощью множества эталонных представителей каждого образа, сохраняющему основные закономерности, необходимые для хорошего распознавания как объектов исходной выборки, так и новых объектов. Такие эталонные объекты будем называть *столпами*. Чем сложнее структура образов, чем сильнее они пересекаются, тем больше столпов потребуется для описания таких данных. Если удастся построить такое описание данных и перейти от образов A и B к множествам столпов этих образов S_A и S_B, соответст-

венно, то вычислять конкурентное сходство объекта z с образом A в конкуренции с образом B можно как $F(z, S_A|S_B)$. Вычисление конкурентного сходства не по всей выборке, а по её сжатому описанию позволяет адаптировать данную меру к особенностям решаемой задачи.

Для построения сжатого описания данных в виде системы столпов используется алгоритм FRiS-Stolp [13]. Алгоритм работает при любом соотношении количества объектов к количеству признаков и при произвольном виде распределения образов. В качестве столпов выбираются объекты, которые обладают высокими значениями двух свойств: *обороноспособности* по отношению к объектам своего образа и *толерантности* по отношению к объектам других образов. Чем выше обороноспособность эталона, тем меньше будет ошибок первого рода (пропуск цели). Чем выше толерантность эталона, тем меньше будет ошибок второго рода (ложная тревога). Набор столпов считается достаточным для описания выборки, если сходство F всех объектов обучающей выборки с ближайшими своими столпами в конкуренции с ближайшими объектами других образов превышает пороговое значение F^*, например, $F^* = 0$.

Отметим некоторые особенности алгоритма FRiS-Stolp. Вне зависимости от вида распределения обучающей выборки столпами выбираются объекты, расположенные в центрах локальных сгустков и защищающие максимально возможное количество объектов с заданной надёжностью. При нормальных распределениях столпами в первую очередь будут выбраны объекты, ближайшие к точкам математического ожидания. Следовательно, при приближении закона распределения к нормальному, решение задачи построения решающих функций стремится к статистически оптимальному. Если распределения полимодальны и образы линейно неразделимы, столпы будут стоять в центрах мод. Сжатое описание образов через множество столпов также можно использовать для распознавания новых объектов.

2 FRiS-компактность. Алгоритм FRiS-GRAD

Практически все алгоритмы распознавания в том или ином виде основаны на использовании гипотезы компактности. При этом в зависимости от модели образы признаются компактными при выполнении одного из нижеперечисленных условий: если они имеют простую форму, разделяются границей простой формы, объекты одного образа похожи друг на друга и не похожи на объекты других образов. Для получения количественной оценки компактности каждого образа в отдельности и качества (информативности) признакового пространства предлагается использовать описанную выше FRiS-функцию, формализующую представление о компактности образа, в соответствии с которым «внутреннее» сходство его объектов друг с другом велико, а «внешнее» сходство с объектами других образов мало.

Действительно, для произвольного объекта $a \in A$ мера конкурентного сходства этого объекта со своим образом в конкуренции с образом B показывает, насколько этот объект похож на свой образ и не похож на образ B. Если эта величина для всех объектов образа A положительна, то можно считать данный образ компактным. Таким образом, вычисляя среднее значение FRiS-функции по всем объектам образа A, можно оценить компактность данного образа. Если при этом вычислять FRiS-функцию с опорой на столпы, то такая оценка компактности будет автоматически адаптироваться к особенностям данных.

Более формально процедура вычисления компактности для случая двух образов выглядит следующим образом.

1) С помощью алгоритма FRiS-Stolp строятся множества столпов S_A и S_B образов A и B.
2) Для каждого элемента $a \in A$ оценивается сходство с ближайшим столпом из множества S_A в конкуренции с ближайшим столпом из S_B. Затем вычисляется FRiS-компактность образа A в конкуренции с образом B:

$$C_{A|B} = \frac{1}{|S_A||A|}\left(\sum_{a\in A} F(a, S_A \mid S_B) - |S_A|\right).$$

Отметим, что количество столпов образа зависит от структуры распределения объектов и величины порога *F**: с ростом *F** увеличивается как число столпов, так и точность описания распределения, но растёт и сложность его описания, т. е. множитель $1/|S_A|$ является штрафом за структурную сложность образа.

3) Аналогично вычисляется величина $C_{B|A}$ FRiS-компактности образа *B* в конкуренции с *A*.
4) Далее получим оценку компактности образов *A* и *B* как усреднение величин $C_{A|B}$ и $C_{B|A}$:

$$C = C_{A|B} + C_{B|A}.$$

Умение оценивать компактность выборки полезно для нахождения информативной подсистемы признаков, так как величина компактности может использоваться как критерий качества набора признаков. В настоящее время преобладают задачи, в которых количество признаков *N* на порядки превышает количество объектов *M*. При этом информация, полезная для решения конкретной классификационной задачи, обычно представлена в нескольких признаках $n << N$. Выбор этих *n* признаков позволяет в дальнейшем не только существенно сократить затраты машинных ресурсов, но и повышает надёжность распознавания образов. Признаки могут зависеть друг от друга, что не позволяет по оценкам индивидуальной информативности каждого признака выбрать подмножество в виде списка из *n* наиболее информативных признаков. Если *n* задано, точное решение можно получить, проверив все сочетания из *N* признаков по *n*, что в реальных задачах часто практически невозможно. По этой причине используются эвристические алгоритмы направленного перебора, например, алгоритмы AdDel [7] или GRAD [8].

Величина FRiS-компактности используется в качестве критерия информативности в алгоритме выбора признаков FRiS-GRAD [14]. Данный критерий применим к любому виду распределений и любому соотношению *M* и *N*. Так как для вычисления FRiS-компактности требуется строить систему столпов, описывающую выборку в каждом из рассматриваемых подпространств, то можно считать, что алгоритм FRiS-GRAD формирует не только набор информативных признаков, но и решающее правило, представленное в виде системы столпов.

Для оценки эффективности алгоритма FRiS-GRAD проводилось его масштабное тестирование на девяти медицинских задачах, объектами в которых выступали пациенты с различными заболеваниями, а признаками — экспрессия генов, полученная с помощью микрочипирования. Особенностью этих задач была их плохая обусловленность, а именно: число признаков на несколько порядков превышало число объектов в выборке.

Результаты работы алгоритма сравнивались с более ранними результатами, полученными четырьмя наиболее часто используемыми алгоритмами распознавания (метод опорных векторов, межгрупповой анализ, байесовский классификатор и метод *k* ближайших соседей) в информативных подпространствах, выбранных десятью известными алгоритмами выбора признаков.

Для оценки качества работы алгоритмов использовалась перекрёстная проверка: 50% выборки использовалось для обучения, а на оставшихся 50% оценивалась надёжность распознавания. Все результаты (кроме относящихся к FRiS-GRAD) были взяты из [15], где для каждой задачи приводилось 40 различных вариантов решений, полученных всеми возможными сочетаниями алгоритмов. Для сравнения были выбраны лучшие результаты по каждой задаче. Результаты сравнения представлены в таблице 2. Здесь показаны: имя задачи, размерность признакового пространства *N*, отношение количества объектов первого образа *M*1 к количеству объектов второго образа *M*2 и два столбца результатов - ожидаемая надёжность

распознавания для лучшего сочетания алгоритмов и ожидаемая надёжность распознавания для алгоритма FRiS-GRAD.

Таблица 2 - Результаты решения девяти задач

Задачи	*N*	*M*1/*M*2	Рекорды	FRiS-GRAD
ALL1	12625	95/33	100.0	100.0
ALL2	12625	24/101	78.2	93.7
ALL3	12625	65/35	59.1	75.3
ALL4	12625	26/67	82.1	93.7
Prostate	12625	50/53	90.2	94.1
Myeloma	12625	36/137	82.9	93.9
ALL/AML	7129	47/25	95.9	100.0
DLBCL	7129	58/19	94.3	96.0
Colon	2000	22/40	88.6	94.2
Average			85.7	93.4

Анализ представленных результатов показывает уверенное превосходство алгоритма FRiS-GRAD перед наиболее популярными алгоритмами анализа данных при решении сложных, плохо обусловленных задач, в которых количество признаков на порядки превосходит количество объектов в выборке.

3 Оценка разделимости классов. Алгоритм FRiS-Censor

Алгоритм FRiS-Stolp наращивает количество столпов до тех пор, пока все объекты не станут надёжно защищенными. Надёжно защищённым считается объект, сходство которого с ближайшим столпом своего образа в конкуренции со столпами образа-конкурента превышает заданный порог *F**. Каждый столп с множеством надёжно защищаемых им объектов образует кластер. Но на последних шагах работы алгоритма возникают столпы, которые защищают малые количества объектов, вплоть до того, что они защищают только самих себя. На роль таких столпов могут попасть «шумовые» объекты (выбросы), свойства которых сильно отличаются от свойств остальных объектов образа. Иногда это говорит об уникальных свойствах таких объектов, однако, более часто причина отличий состоит во влиянии неучитываемых факторов, таких как сбой измерительных приборов, ошибки занесения данных в протокол и пр. Встречаются и объекты, которые не являются «ошибочными», но находятся на периферии распределения и оказываются глубоко в зоне пересечения с соседним образом. Они тоже могут неоправданно сильно усложнить решающие правила, делая их «вычурными». Таким образом, «погоня» за мелкими и единичными кластерами ведёт к переобучению.

Качество описания обучающей выборки (или оценка разделимости классов) зависит от набора выбранных эталонов. В случае, когда каждый объект выборки является столпом, информация о ней сохраняется полностью, но для распознавания такое описание не пригодно. О том, насколько хорошо некоторая система столпов описывает классы в заданном признаковом пространстве, будем судить по оценке качества описания обучающей выборки, опирающейся на понятие компактности.

В случае двух классов для каждого элемента обучающей выборки оценивается его сходство с ближайшим столпом своего образа в конкуренции с ближайшим объектом из конкурирующего образа и затем вычисляется качество описания обучающей выборки *L* столпами:

$$H(L) = \frac{1}{|S_A \cup S_B \| A \cup B|}\left(\sum_{a\in A} F(a, S_A \mid B) + \sum_{b\in B} F(b, S_B \mid A)\right).$$

При изменении количества столпов L меняется качество описания H обучающей выборки и ошибка распознавания E тестовой выборки. Выдвигается и проверяется гипотеза о том, что между функциями $H(L)$ и $E(L)$ имеется закономерная связь, используя которую можно найти такое количество столпов, что дальнейшее увеличение числа столпов ведёт к переобучению.

Проверка этой гипотезы проводилась на модельной задаче распознавания двух образов, каждый из которых представлял собой суперпозицию нескольких (от 2-х до 4-х) нормально распределённых кластеров в двумерном пространстве признаков. Рассматривалось 10 распределений, которые отличались друг от друга количеством образующих нормальных компонентов, их дисперсиями, координатами математических ожиданий и количеством объектов в компонентах. Каждый образ был представлен 250 объектами. При каждом распределении выборка 100 раз случайным способом делилась на две части: обучающую (по 50 объектов первого и второго образов) и контрольную (по 200 объектов каждого образа). Количество экспериментов при различных численных реализациях исходных данных было равно 1000.

Результаты отдельных экспериментов, приведенные на рисунке 2, служат подтверждением выдвинутого предположения. Таким образом, сформулирована и экспериментально подтверждена гипотеза о том, что точка перегиба кривой (первый локальный максимум функции H), описывающей разделимость классов, может служить сигналом о начале переобучения. Объекты, оставшиеся незащищёнными, считались выбросами и исключались из рассмотрения.

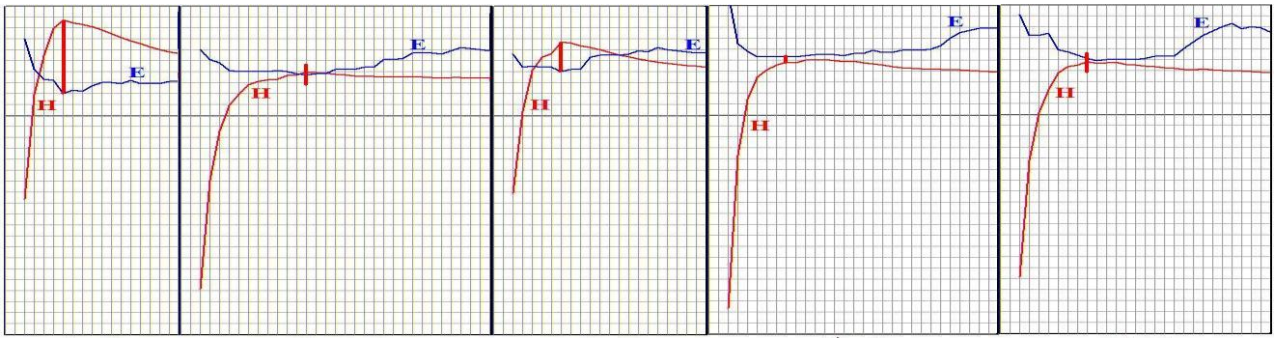

Рисунок 2 - Графики качества описания обучающей выборки (H) и графики ошибки распознавания (E) в зависимости от числа выбранных эталонов

Алгоритм FRiS-Censor [16], отыскивающий эту точку перегиба, позволил улучшить качество распознавания в сравнении со стандартным алгоритмом FRiS-Stolp на 3-5%. При этом число столпов, достаточных для описания выборки, снижалось в среднем в 3 раза относительно числа столпов, достаточных для защиты всех объектов выборки.

Заключение

Функция конкурентного сходства оказалась универсальным инструментом для разработки алгоритмов решения различных типов задач когнитивного анализа данных. Наиболее интересные и убедительные результаты работоспособности этой модели были получены для задачи выбора информативной системы признаков в случае плохо обусловленных задач сложной структуры. Помимо задач, описанных выше, с её помощью удалось решить задачу таксономии (алгоритм FRiS-Tax [17]), задачу частичного обучения (алгоритм FRiS-TDR [18]), задачу заполнения пробелов в двух- и трёхмерных таблицах данных (FRiS-ZET и 3D-ZET [19]). Все эти алгоритмы показали свою эффективность и активно используются для решения прикладных задач в самых разных областях.

Достижение таких результатов было бы невозможным без светлых идей и чуткого руководства Загоруйко Николая Григорьевича. Мы благодарны судьбе за возможность работать с этим удивительно талантливым и мудрым человеком. ***Светлая ему память.***

Благодарности

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, проект № 14-01-00039.

Список источников

[1] ***Айвазян, С.А.*** Прикладная статистика. Основы моделирования и первичная обработка данных / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1983. – 471 с.

[2] ***Arkadev, A.G.*** Computers and Pattern Recognition / A.G. Arkadev, E.M. Braverman. - Thompson Book Company, 1967. – 115 p.

[3] ***Vapnik, V.N.*** Statistical Learning Theory / V.N. Vapnik. – NY: Wiley-Interscience, 1998. – 740 p.

[4] ***Журавлев, Ю.И.*** Об алгебраическом подходе к решению задач распознавания или классификации / Ю.И. Журавлев // Проблемы кибернетики. – 1978. – Т. 33. – С. 5-68.

[5] ***Ivakhnenko, A.G.*** Cybernetics and forecasting techniques / A.G. Ivakhnenko, V.G. Lapa. – NY.: Elsevir Publishing Company, 1967. – 168 p.

[6] ***Шлезингер, М.И.*** Математические средства обработки изображений / М.И. Шлезингер. - К.: Наукова думка, 1989. – 198 с.

[7] ***Загоруйко, Н.Г.*** Когнитивный анализ данных / Н.Г. Загоруйко. – Новосибирск: Академическое изд-во ГЕО, 2013. – 186 с.

[8] ***Загоруйко, Н.Г.*** Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Изд-во ИМ СО РАН, 1999. – 270 с.

[9] ***Загоруйко, Н.Г.*** Методы распознавания и их применение / Н.Г. Загоруйко. – М.: Советское радио, 1972. – 208 с.

[10] ***Zagoruiko, N.G.*** A quantitative measure of compactness and similarity in a competitive space / N.G. Zagoruiko, I.A. Borisova, V.V. Dyubanov, O.A. Kutnenko // Journal of Applied and Industrial Mathematics. – 2011. - V. 5. - №1. – P. 144-154.

[11] ***Kira, K.*** The Feature Selection Problem: Traditional Methods and a New Algorithm / K. Kira, L. Rendell // Proc. 10 Nat. Conf. Artificial Intelligence (AAAI-92). - Menlo Park: AAAI Press, 1992. – P. 129-134.

[12] ***Rousseeuw, P.J***. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis / P.J. Rousseeuw // J. Comput. Appl. Math. – 1987. - V. 20. – P. 53–65.

[13] ***Zagoruiko, N.G.*** A construction of a compressed description of data using a function of rival similarity / N.G. Zagoruiko, I.A. Borisova, V.V. Dyubanov, O.A. Kutnenko // Journal of Applied and Industrial Mathematics. – 2013. - V. 7. - № 2. – P. 275-286.

[14] ***Zagoruiko, N.*** Attribute selection through decision rule construction (algorithm FRiS-GRAD) / N. Zagoruiko, I. Borisova, V. Dyubanov, O. Kutnenko // Proc. of 9 Intern. Conf. Pattern recognition and Image Analysis (PRIA-2008). – Nizhniy Novgorod. 2008. V. 2. – P. 335-338.

[15] ***Jeffery, I.*** Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data / I. Jeffery, D. Higgins, A. Culhane // BMC Bioinformatics. – 2006. – DOI:10.1186/1471-2105-7-359. - http://www.biomedcentral.com/1471-2105/7/359. (актуально на 03.03.2015).

[16] ***Загоруйко, Н.Г.*** Обучение распознаванию без переобучения / Н.Г. Загоруйко, О.А. Кутненко, А.О. Зырянов, Д.А. Леванов // Машинное обучение и анализ данных. – 2014. - Т. 1. - №7. – С. 891-901.

[17] ***Борисова, И.А.*** Алгоритм таксономии FRiS-Tax / И.А. Борисова // Научный вестник НГТУ – Новосибирск: Изд-во НГТУ, 2007. - №3. – С. 3-12.

[18] ***Borisova, I.A.*** Feature selection by using the FRiS-function in the task of generalized classification / I.A. Borisova, N.G. Zagoruiko // Pattern Recognition and Image Analysis. – 2011. -V. 21. - №2. – P. 117-120.

[19] ***Zagoruiko, N.G.*** Error detection and gap filling in cubes of data / N.G. Zagoruiko, V.V. Tatarnikov // Journal of Applied and Industrial Mathematics. – 2014. - V. 8. - Issue 3. – P. 444-451.

RIVAL SIMILARITY AS AN UNIVERSAL BASIC TOOL OF COGNITIVE DATA MINING

N.G. Zagoruiko, I.A. Borisova[1], O.A. Kutenko[2], V.V. Dyubanov[3], D.A. Levanov[4]

Sobolev Institute of Mathematics, SB RAS
Design Technological Institute of Digital Techniques, SB RAS
Novosibirsk State University, Novosibirsk, Russia
[1]*biamia@mail.ru,* [2]*olga@math.nsc.ru*

Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
[3]*vladimir.dyubanov@gmail.com,* [4]*levanovd@gmail.com*

Abstract

During Data Mining tasks solving a person use a specific universal psycho and physiological cognitive technique. Key points of the technique are in a way of estimating a measure of similarity between objects and in necessity to maximize compactness and simplicity of a world description according to the measure. Nikolay Zagoruiko offered a measure of similarity, which takes into account a rival environment. In the paper the Function of Rival Similarity (FRiS-function) and some possibilities of its usage for patterns compactness and separability estimating are presented. These estimations are used in algorithms for solving classification (FRiS-Stolp algorithm), feature selection (FRiS-GRAD algorithm) and censoring (FRiS-Censor algorithm) tasks. Main ideas, some properties of the algorithms and their results on model and real tasks of data mining are described in the paper as well.
Key words: *cognitive data mining, pattern recognition, feature selection, censoring, function of rival similarity.*

Acknowledgment

This work was supported by the Russian Foundation for Basic Research, investigations, project № 14-01-00039.

References

[1] ***Aivazian, S.A.*** Prikladnaja statistika. Osnovy modelirovanija i pervichnaja obrabotka dannyh [Applied statistics. Essential principles of modeling and data preprocessing / S.A. Aivazian, I.S. Enyukov, L.D. Meshalkin. – Moscow: Finansy i statistika, 1983. – 471 p. (In Russian).

[2] ***Arkadev, A.G.*** Computers and Pattern Recognition / A.G. Arkadev, E.M. Braverman. - Thompson Book Company, 1967. – 115 p.

[3] ***Vapnik, V.N.*** Statistical Learning Theory / V.N. Vapnik. – NY: Wiley-Interscience, 1998. – 740 p.

[4] ***Zhyravlev, U.I.*** Ob algebraicheskom podhode k resheniju zadach raspoznavanija ili klassifikacii [About algebraic approach to solving tasks of recognition and classification] / U.I. Zhyravlev // Problemy kibernetiki. – 1978. – V. 33. – P. 5-68. (In Russian).

[5] ***Ivakhnenko, A.G.*** Cybernetics and forecasting techniques / A.G. Ivakhnenko, V.G. Lapa. – NY.: Elsevir Publishing Company, 1967. – 168 p.

[6] ***Shlezinger, M.I.*** Matematicheskie sredstva obrabotki izobrazhenij [Math-based environment for image processing] / M.I. Shlezinger. – Kiev: Naukova dumka, 1989. – 198 p. (In Russian).

[7] ***Zagoruiko, N.G.*** Kognitivnyj analiz dannyh [Cognitive data mining] / N.G. Zagoruiko. – Novosibirsk: Academic published house GEO, 2013. – 186 p. (In Russian).

[8] ***Zagoruiko, N.G.*** Prikladnye metody analiza dannyh i znanij [Applied methods of data and knowledge mining] / N.G. Zagoruiko. – Novosibirsk: Published house IM SB RAS, 1999. – 270 p. (In Russian).

[9] ***Zagoruiko, N.G.*** Metody raspozanvanija i ih primenenie [Methods of pattern recognition and their applications] / N.G. Zagoruiko. – Moscow: Sovetskoe radio, 1972. – 208 p. (In Russian).

[10] ***Zagoruiko, N.G.*** A quantitative measure of compactness and similarity in a competitive space / N.G. Zagoruiko, I.A. Borisova, V.V. Dyubanov, O.A. Kutnenko // Journal of Applied and Industrial Mathematics. – 2011. - V. 5. - №1. – P. 144-154.

[11] ***Kira, K.*** The Feature Selection Problem: Traditional Methods and a New Algorithm / K. Kira, L. Rendell // Proc. 10 Nat. Conf. Artificial Intelligence (AAAI-92). - Menlo Park: AAAI Press, 1992. – P. 129-134.
[12] ***Rousseeuw, P.J***. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis / P.J. Rousseeuw // J. Comput. Appl. Math. – 1987. - V. 20. – P. 53–65.
[13] ***Zagoruiko, N.G.*** A construction of a compressed description of data using a function of rival similarity / N.G. Zagoruiko, I.A. Borisova, V.V. Dyubanov, O.A. Kutnenko // Journal of Applied and Industrial Mathematics. – 2013. - V. 7. - № 2. – P. 275-286.
[14] ***Zagoruiko, N.*** Attribute selection through decision rule construction (algorithm FRiS-GRAD) / N. Zagoruiko, I. Borisova, V. Dyubanov, O. Kutnenko // Proc. of 9 Intern. Conf. Pattern recognition and Image Analysis (PRIA-2008). – Nizhniy Novgorod. 2008. V. 2. – P. 335-338.
[15] ***Jeffery, I.*** Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data / I. Jeffery, D. Higgins, A. Culhane // BMC Bioinformatics. – 2006. – DOI:10.1186/1471-2105-7-359. - http://www.biomedcentral.com/1471-2105/7/359. (valid on 03.03.2015).
[16] ***Zagoruiko, N.G.*** Obuchenie raspoznavaniju bez pereobuchenija / N.G. Zagoruiko, O.A. Kutnenko, A.O. Zyrjanov, D.A. Levanov // Mashinnoe obuchenie i analiz dannyh [Machine learning and data mining] – 2014. - V. 1. - №7. – P. 891-901. (In Russian).
[17] ***Borisova, I.A.*** Algoritm taksonomii FRiS-Tax [FRiS-Tax algorithm of taxonomy] / I.A. Borisova // Nauchnyj vestnik NGTU– Novosibirsk: NSTU Publ., 2007. - №3. – P. 3-12. (In Russian).
[18] ***Borisova, I.A.*** Feature selection by using the FRiS-function in the task of generalized classification / I.A. Borisova, N.G. Zagoruiko // Pattern Recognition and Image Analysis. – 2011. -V. 21. - №2. – P. 117-120.
[19] ***Zagoruiko, N.G.*** Error detection and gap filling in cubes of data / N.G. Zagoruiko, V.V. Tatarnikov // Journal of Applied and Industrial Mathematics. – 2014. - V. 8. - Issue 3. – P. 444-451.

Сведения об авторах

Загоруйко Николай Григорьевич, (1931-2015). В 1953 году окончил с отличием электротехнический факультет Ленинградского института киноинженеров (ЛИКИ), получив диплом инженера-электрика по специальности «звукотехника». Работал в ЛИКИ, исследуя проблемы магнитной записи сигналов в запоминающих устройствах ЭВМ, в 1960 г. перешёл в Институт математики СО АН СССР. В 1962 году защитил кандидатскую диссертацию. Впоследствии занимался проблемами технической кибернетики, связанными с методами автоматического распознавания образов. В 1969 году защитил докторскую диссертацию.

Круг вопросов, рассматриваемых в 240 опубликованных им работах, достаточно широк. Вначале были работы, связанные с расчётами пространственных полей магнитных головок. Затем последовала большая серия работ по проблемам обмена устной информацией между человеком и машиной. В этой области созданная им в 1962 году лаборатория стала одним из ведущих научных коллективов в СССР. По инициативе Н.Г. Загоруйко в 1963 году была организована Всесоюзная школа-семинар по проблеме «Автоматическое распознавание слуховых образов» (АРСО), которая регулярно проводилась в течение почти 30 лет и объединяла в неформальный коллектив все ведущие научные организации СССР, связанные с исследованием речевых сигналов. Н.Г. Загоруйко был постоянным председателем Программного комитета АРСО, которая собиралась 17 раз в различных городах Союза. Тесное сотрудничество математиков, инженеров, психологов, акустиков и лингвистов позволяло советским учёным занимать ведущие мировые позиции в области распознавания и синтеза речи.

Занятия распознаванием речи естественным путем переросли в исследования задач распознавания образов и более общей проблемы автоматического обнаружения эмпирических закономерностей. Разрабатываемые Н.Г. Загоруйко и его сотрудниками методы автоматической классификации (таксономии), выбора информативных признаков, построения решающих функций, прогнозирования, обнаружения ошибок и заполнения пробелов в таблицах данных получили широкое применение в геологии, медицине, генетике, экономике, гидроакустике и во многих других прикладных областях.

В течение 2-х лет Н.Г. Загоруйко возглавлял группу исследователей в Международной лаборатории искусственного интеллекта в Братиславе, читал лекции в ряде зарубежных университетов (Королевский технологический институт, Швеция; Южно-Калифорнийский университет, США; Вроцлавский университет, Польша;

Киотский университет, Япония и др.). Н.Г. Загоруйко работал профессором НГУ с 1969 года, в течение 8 лет был проректором НГУ по научной работе. Среди его учеников 22 кандидата и 6 докторов наук.

С 2007 года был сопредседателем Программного комитета конференции «Знания-Онтологии-Теории» (ЗОНТ), с 2011 года - членом редколлегии журнала «**Онтология проектирования**».

Активная жизненная позиция отличала Николая Григорьевича не только в науке. Он был одним из организаторов клуба межнаучного общения «Под интегралом», директором Молодежного научно-производственного объединения «Факел», одним из инициаторов антиалкогольного общественного движения в Советском Союзе. Он являлся мастером спорта СССР и судьей Всесоюзной категории по современному пятиборью, был солистом ансамбля ЛИКИ.

Н.Г. Загоруйко был награжден Орденом Знака Почёта, двумя серебряными медалями ВДНХ СССР.

Nikolay Grigorevich Zagoruiko (1931-2015) graduated from the Leningrad Institute of Motion-picture Engineers in 1953. Received candidate's (in 1962) and doctoral (in 1969) degrees in pattern recognition. In 1988–1990 had been heading a project in the International Laboratory of Artificial Intelligence in Bratislava (Slovakia). He was a chief researcher at the Laboratory of Data Mining at the Institute of Mathematics, Siberian Branch, Russian Academy of Sciences. He was Professor at Novosibirsk State University. He was co-author more 200 scientific articles and monographs in the field of DM and AI.

Борисова Ирина Артемовна, 1978 г. рождения. Окончила Новосибирский государственный университет в 2002 г., к.т.н. (2008). Старший научный сотрудник лаборатории анализа данных Института математики им. С.Л. Соболева СО РАН. В списке научных трудов более 60 статей, в области анализа данных и распознавания образов.

Irina Artemovna Borisova (b. 1978) graduated from the Novosibirsk State University in 2002, C. Sc. Eng. (2008). She is a senior researcher of the Laboratory of Data Mining at the Institute of Mathematics, Siberian Branch, Russian Academy of Sciences. She is a co-author of more than 60 publications in the field of Data Mining and Pattern recognition.

Кутненко Ольга Андреевна. Родилась в 1958 г. Окончила в 1980 г. Новосибирский государственный университет. В 2000 г. защитила кандидатскую диссертацию. Старший научный сотрудник лаборатории анализа данных Института математики им. С.Л. Соболева СО РАН. Автор более 70 научных работ. Научные интересы: анализ данных, обнаружение эмпирических закономерностей, распознавание образов.

Olga Andreevna Kytnenko (b. 1958) graduated from the Novosibirsk State University in 1980, C. Sc. Eng. (2000). She is senior researcher of the Laboratory of Data Mining at the Institute of Mathematics, Siberian Branch, Russian Academy of Sciences. She is co-author of more than 70 publications in the field of Data Mining and Pattern recognition.

Дюбанов Владимир Владимирович, родился в 1981 г. Окончил магистратуру Новосибирского государственного университета в 2004 г. С 2004 года является сотрудником лаборатории анализа данных Института математики им. С.Л. Соболева СО РАН. Соавтор 17 статей. Научные интересы: анализ данных, машинное обучение.

Vladimir Vladimirovich Dyubanov (b. 1981) graduated from magistrate of Novosibirsk State Universiy in 2004. Since 2004 had worked at Laboratory of Data Mining at the Institute of Mathematics, Siberian Branch, Russian Academy of Sciences. Co-author of 17 articles. Research interests include data analysis, machine learning.

Леванов Дмитрий Александрович. Родился в 1989 г. Окончил в 2012 г. Новосибирский государственный университет. Аспирант лаборатории анализа данных Института математики им. С.Л. Соболева СО РАН. Автор 3 статей. Научные интересы: анализ данных, обнаружение эмпирических закономерностей, распознавание образов.

Levanov Dmitry Alexandrovich, (b. 1989) graduated from Novosibirsk State University in 2012. Post-graduate at Laboratory of Data Mining at the Institute of Mathematics, Siberian Branch, Russian Academy of Sciences. Author of three articles. Research interests include data analysis, information retrieval and image recognition.