КРИТЕРИИ ИНФОРМАТИВНОСТИ И ПРИГОДНОСТИ ПОДМНОЖЕСТВА ПРИЗНАКОВ

Ирина Борисова, Николай Загоруйко, Ольга Кутненко

***Аннотация**: Предлагается вариант решения проблемы Колмогорова о выборе подсистемы признаков, которая была бы не только информативна, но и пригодна для распознавания контрольной выборки. Способ решения состоит в использовании нового критерия информативности признаков в виде функции сходства и различия.*

***Ключевые слова**: Распознавание образов, информативность, пригодность, функция сходства и различия, компактность.*

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

Введение

В 1933 году А.Н. Колмогоров опубликовал работу [1], в которой обратил внимание на трудности, связанные с решением проблемы выбора подмножества информативных предикторов при построении регрессионных уравнений для случая, когда количество потенциальных предикторов сравнимо или превышает количество наблюдаемых объектов. Если предикторы зависят друг от друга, то выбор наиболее информативного подмножества из их большого исходного количества представляет собой NP-трудную переборную задачу. Но дело не только в этом. Встречаются задачи, в которых основная часть характеристик не имеет прямого отношения к целевой функции и потому играет роль случайного шума. Чем больше таких характеристик и чем меньше наблюдаемых объектов, тем выше вероятность обнаружения «псевдоинформативного» набора из шумовых предикторов.

В последние годы актуальность проблемы выбора информативного подмножества признаков и оценки его пригодности для решения задач регрессионного анализа и распознавания образов сильно возросла. Стали встречаться реальные задачи распознавания образов, например, в генетике, в которых небольшое число (десятки) объектов обучающей выборки описывается очень большим числом характеристик (десятками тысяч).

Успех в решении этой проблемы зависит от того, как организована **процедура** направленного перебора вариантов, по каким **критериям** оценивается **информативность** и **пригодность** конкурирующих вариантов подсистем признаков. Первая составляющая успеха в решении задачи выбора подсистем признаков претерпела в последние годы заметное развитие. В данной работе рассматриваются вторая и третья часть проблемы Колмогорова, связанные с критериями информативности и неслучайности подсистем-конкурентов.

Вероятность случайного выбора

Если объем обучающей выборки (M) мал, а количество исходных признаков (N) велико, то вероятность того, что в состав информативного подмножества из n<N признаков могут попасть случайные признаки. Ясно, что эта вероятность будет увеличиваться с ростом размерности выбираемого подпространства n и отношения N/M. Для оценки характера зависимости вероятности случайного результата от параметров

N, M и n был проделан машинный эксперимент с таблицами случайных чисел. Количество объектов было равным 75, а размерность таблицы менялся от 10 до 2000. Объекты случайным образом делились на два класса (по 50% объектов в каждом образе). В каждой такой таблице методом AdDel [2] выбирались подсистемы из наиболее информативных признаков. Количество признаков в подсистемах менялось от 1 до 22. На рис. 1 показаны усредненные по 10 экспериментам результаты распознавания обучающей выборки в режиме скользящего экзамена по выбранным признакам.

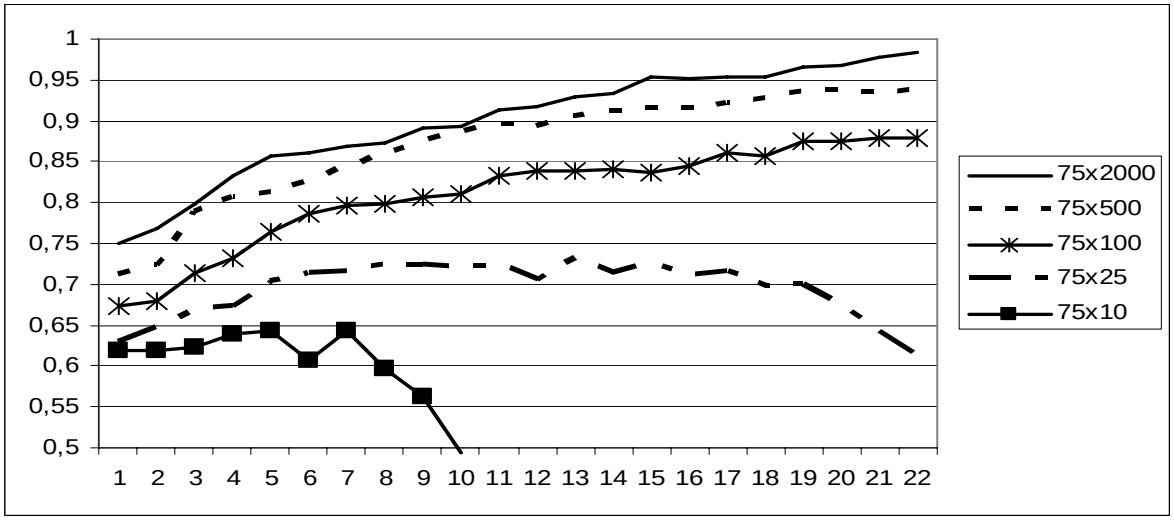

Рис. 1. Вероятность правильного распознавания случайной обучающей выборки для таблиц с параметрами $M = 75$ и N от 2000 до 10 при n от 1 до 22..

Из результатов видно, что при больших N можно найти случайное сочетание n случайных признаков, которые на обучающей выборке покажут свою высокую информативность.

Обратим теперь внимание на следующий вопрос: какой критерий информативности признаков будет защищать нас от случайного выбора наиболее эффективно?

Функция сходства и различия

В описанных экспериментах, как и в большинстве существующих методов, оценкой информативности U подсистем на этапе обучения служило количество правильно распознанных объектов обучающей выборки в режиме скользящего экзамена. При этом решающее правило, по которому контрольный объект z относился к первому образу, было основано на сравнении расстояний r от контрольного объекта z до эталонов первого (r1) и второго (r2) образов. Зная эти расстояния, можно использовать простое правило ближайшего соседа (kNN): если r1<r2, то объект z принадлежит первому образу, и наоборот. Но оказалось, что знанием величин r1 и r2 можно воспользоваться и более эффективно, если ввести следующие функции:

$$F_1 = (r_2 - r_1)/(r_1 + r_2) \text{ и } F_2 = (r_1 - r_2)/(r_1 + r_2)$$

Значения этих функций меняются в пределах от +1 до -1, а их сумма всегда равна 0. Если контрольный объект z совпадает с эталоном первого образа, то r1=0 и F1=1, а F2= -1. Это говорит об абсолютном сходстве объекта z с эталоном первого образа и о максимальном его отличии от эталона второго образа. При расстояниях r1=r2 значения F1=F2=0, что указывает на границу между образами. В точках границы

объект в равной степени похож и не похож на эти конкурирующие образы. Функция F хорошо согласуется с механизмами восприятия сходства и различия, которыми пользуется человек, сравнивая некий объект с двумя другими объектами. Мы будем называть F функцией сходства и различия (ФСР). Функция ФСР применима для решения многих задач анализа данных: для автоматической классификации, построения решающих правил и других. Оказалась она полезной и в качестве критерия информативности признаков. Если, например, объекты двух образов представлены двумя линейно разделимыми группами объектов, то оценка информативности, найденная по критерию U числа правильно распознаваемых объектов, не будет зависеть от расстояния между группами. А среднее значение функции сходства и различия (Fcp) будет зависеть от того, как близко группы находятся от разделяющей границы. Те объекты, которые располагаются в тесном окружении своих объектов и значительно удалены от объектов других образов, имеют более высокое значение функции F, чем периферийные объекты, близкие к другим образам.

Возникла идея сравнить между собой два критерия информативности – число правильно распознанных объектов обучающей выборки (U) и среднее значение функции принадлежности (F_S). Третий критерий вытекает из предложения Фишера оценивать информативность признаковых систем по величине, пропорциональной расстоянию между математическими ожиданиями образов, деленному на сумму их дисперсий: $Q = |\mu_1 - \mu_2| / (\sigma_1 - \sigma_2)$.

При высокой размерности признакового пространства и малом количестве обучающих объектов можно в качестве аналога математического ожидания образа использовать координаты центра тяжести его объектов, а в качестве дисперсий – среднее расстояние между объектами образа.

Эти три критерия – U, F_S и Q - сравнивались в следующем модельном эксперименте. Исходные данные состояли из 200 объектов двух образов (по 100 объектов каждого образа) в 100-мерном пространстве. Признаки генерировались так, чтобы они обладали разной информативностью. В итоге около 30 признаков оказывались в той или иной степени информативными, а остальные признаки генерировались датчиком случайных чисел и были заведомо неинформативными. По этой таблице алгоритмом AdDel выбирались наиболее информативные подсистемы размерности n (от 1 до 22). При этом для обучения случайно выбиралось по 35 объектов каждого образа. На контроль предъявлялись остальные 130 объектов.

Надежности распознавания контрольной выборки при использовании критериев U, F_S и Q, усредненные по 10 экспериментам, показаны на рис. 2. Из них видно, что признаки, выбранные по критерию Q, лучше выбранных по критерию ошибок U, но хуже выбранных по функции принадлежности F_S. Это можно объяснить тем, что меры Q и F_S меньше зависят от характеристик отдельных пограничных объектов, чем мера U. В свою очередь, мера Фишера Q ориентирована на разделение нормальных распределений с помощью линейных решающих функций, в то время, как мера F_S адаптируется к особенностям распределения обучающей выборки и соответствует более мощной кусочно-линейной разделяющей границе.

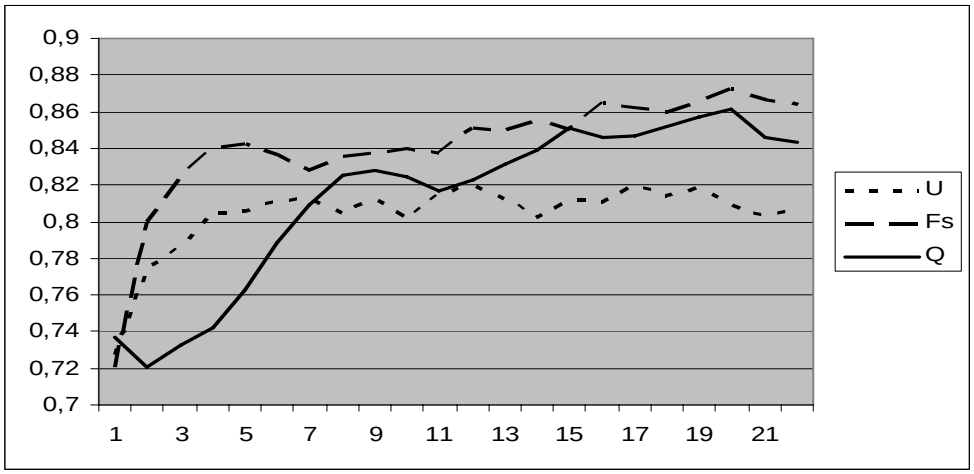

Рис. 2. Результаты выбора подсистем признаков при использовании трех критериев: по числу ошибок (U), по функции принадлежности (F_S) и по критерию Фишера Q.

Критерии U и F_S исследовались на устойчивость к помехам. Для этого исходная таблица из предыдущего эксперимента искажалась шумами разной интенсивности и при каждом уровне шума (от 0,05 до 0,3) выбирались наилучшие подсистемы по этим критериям. Результаты представлены на рисунке 3, из которого видно, что критерий F_S более устойчив, чем критерий U. Результаты на контроле показывают высокую степень корреляции критерия Fs с результатами, полученными на обучении. Это свидетельствует о высоких прогностических свойствах критерия Fs.

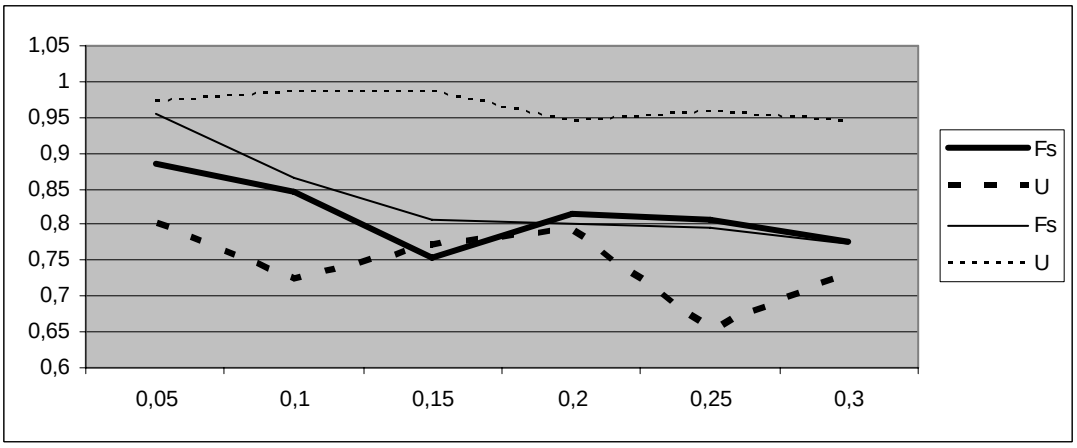

Рис. 3. Результаты обучения и распознавания по критериям U и F_S при разных уровнях шумов. Тонкие линии – обучение, жирные – контроль.

Оценка «пригодности» выбранных подсистем

Для сравнения этих результатов с чисто случайными результатами датчиком случайных чисел с равномерным распределением были сформированы 10 таблиц такого же размера $M = 200$, $N = 100$. Два образа (по 100 объектов) были сформированы методом случайного выбора. По этим чисто шумовым данным для каждой размерности подсистем n выбирались наиболее информативные признаки и определялись значения критерия F_S. Оказалось, что они лежат в «случайном коридоре» с границами от

0,61 до 0,67. Значения F_S для подсистем, найденных по исходной таблице, лежат значительно выше этого коридора и потому могут считаться неслучайными.

Из приведенных результатов можно сформулировать следующую практическую рекомендацию. Определяется значение F_S для наилучшей подсистемы из n^* признаков X, выбранных по обучающей таблице $N \times M$. Затем формируется серия случайных таблиц с такими же значениями N и M, и по ним находятся значения F_S для «лучших» подсистем той же размерности n^*. Если величина F_S для исходной таблицы попадает в пределы значений F_S для случайных таблиц, то можно считать, что выбранные признаки X «псевдоинформативны». Они не пригодны для дальнейшего использования.

По расстоянию между значением критерия F_S подсистемы, выбранной в реальной таблице, и границами «случайного коридора», полученного на наборе случайных таблиц того же размера, можно судить о неслучайности, пригодности выбранных подсистем.

Проверка на реальных данных

Для подтверждения преимуществ критерия F_S перед критерием U на реальных задачах был проведен эксперимент со спектральными данными. Обучающая выборка состояла из двух образов по 25 объектов, выбранных случайно из таблицы реальных спектральных данных двух классов веществ. Из исходного множества 1024 спектральных характеристик формировались два списка из 46 наиболее информативных «вторичных» признаков в виде не перекрывающихся участков спектра. Один список включал в свой состав признаки, отобранные по критерию U, а второй – по критерию F_S. Затем всем n признакам каждого списка в отдельности предъявлялись для распознавания 200 контрольных объектов (по 100 объектов каждого образа). Надежность распознавания по каждому из 46 наиболее информативных признаков представлена на рис. 4.

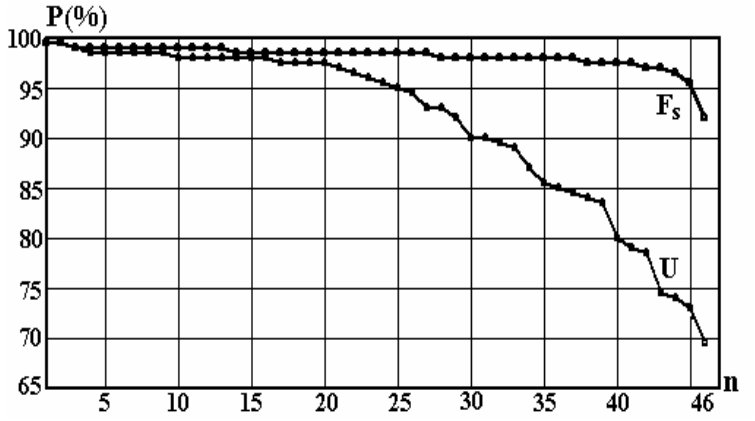

Рис. 4. Надежность P распознавания контрольной выборки по каждому из 46 наиболее информативных признаков, упорядоченных по информативности. Верхняя кривая соответствует выбору по критерию F_S, нижняя – по критерию U.

Эти результаты также подтверждают преимущество выбора признаков по среднему значению функции сходства и различия (F_S) по сравнению с широко распространенным выбором по числу правильно распознанных объектов обучающей выборки (U).

Заключение

Проведенные исследования позволяют сделать следующие выводы:

1. Для оценки информативности признаков или признаковых систем следует использовать не количество правильно распознанных объектов обучающей выборки (U), а среднее значение функции F_S сходства объектов обучающей выборки с эталонами своих образов.

2. Значения меры F_S, получаемые на обучающей таблице и на серии случайных таблиц того же размера, позволяют получить качественную оценку пригодности выбранного подпространства признаков.

3. Значение функции F сходства контрольного объекта с эталоном того или иного образа дает возможность сопроводить результат распознавания оценкой правильности этого результата.

Библиография

1. Колмогоров А.Н. К вопросу о пригодности найденных статистическим путем формул прогноза. - Заводская лаборатория. 1933. №1. С. 164-167.

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

Информация об авторах

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

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

Ольга Кутненко - Институт Математики СО РАН, пр. Коптюга, дом 4, Новосибирск, 630090, Россия; e-mail: olga@math.nsc.ru