Публикация: Труды Всероссийской Конференции «Знания-Онтологии-Теории» (ЗОНТ-07), Новосибирск, 2007, Том I, с.37-44.

Знания-Онтологии-Теории (ЗОНТ-07)

Использование FRiS-функции для построения решающего правила и выбора признаков
(задача комбинированного типа DX)

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

Институт математики им. Соболева СО РАН, пр. Коптюга, д. 4, г. Новосибирск, 630090, Россия.

biamia@mail.ru, vladimir.dyubanov@gmail.com, zag@math.nsc.ru, olga@math.nsc.ru

Аннотация. *Предлагается способ решения задачи комбинированного типа DX – построения решающей функции D в наиболее информативном подпространстве признаков X. При решении этой задачи используется функция конкурентного сходства (FRiS-функция). В итоге выбирается подмножество признаков, в пространстве которых каждый образ представляется таким (необходимым и достаточным) набором эталонов (столпов), которые обеспечивают максимальное значение среднего сходства всех объектов обучающей выборки со своими эталонами. Показывается преимущество критерия информативности, основанного на FRiS-функции, по сравнению с критерием минимума ошибок при скользящем экзамене обучающей выборки. Эта функция упрощает процесс распознавания контрольных объектов, позволяет оценивать надежность распознавания конкретного объекта и предложить вариант ответа на вопрос, поставленный Колмогоровым, о «пригодности» признаков, выбираемых статистическим путем.*

Ключевые слова

Распознавание образов, функция конкурентного сходства, выбор эталонов, выбор признаков, информативность и пригодность признаков

1 Введение

В распознавании образов существует целый класс задач, решение которых стандартными статистическими методами представляется невозможным. Ярким примером такого рода являются задачи, в которых число объектов невелико и меньше числа описывающих их характеристик.

Еще 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. Это происходит потому, что основная часть описывающих 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** различных вариантов подсистем признаков и как устроена непосредственно процедура **распознавания**. Первая (процедурная) составляющая успеха в решении задачи выбора подсистем признаков претерпела в последние годы заметное развитие. В данной работе рассматриваются те части проблемы, которые связаны с критериями информативности и пригодности (неслучайности) признаков и с организацией процедуры распознавания. Для решения этих проблем привлекается новый инструмент в виде функции конкурентного сходства или FRiS-функции. На базе этой функции удалось решить задачу комбинированного типа DX, в которой одновременно строится решающая функция и выбирается наиболее информативное и пригодное подмножество признаков.

Рассмотрим подробнее отдельные части обсуждаемой проблемы.

2 Вероятность случайного выбора

Если объем обучающей выборки (M) мал, а количество исходных признаков (N) велико, то есть 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 M было равным 75, а размерность таблицы N менялась 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 n в подсистемах менялось от 1 до 22. Информативность оценивалась по надежности распознавания обучающей выборки при скользящем экзамене по правилу ближайшего соседа. На рис. 1 показаны результаты, усредненные по 10 экспериментам.

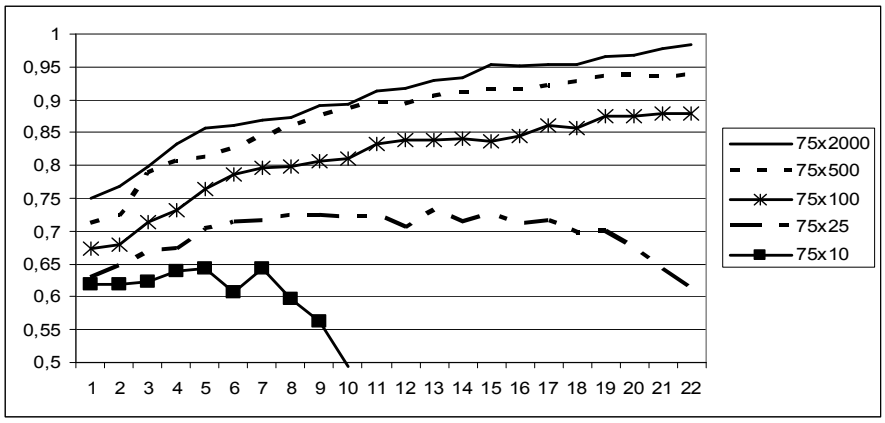

Рис. 1. Вероятность правильного распознавания случайной обучающей выборки для таблиц с параметрами $M = 75$ и N от 2000 до 10 при n от 1 до 22.

Из результатов эксперимента видно, что при больших N можно найти сочетание из n случайных признаков, которые на обучающей выборке покажут свою высокую информативность, но заведомо не пригодны для распознавания контрольной выборки.

Попытаемся ответить на следующий вопрос: какой критерий информативности признаков будет защищать нас от случайного выбора наиболее эффективно?

3 Функция конкурентного сходства

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 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 конкурентного сходства (FRiS-функцией). FRiS-функция 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 конкурентного сходства (Fср) 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 три критерия информативности – долю правильно распознанных объектов обучающей выборки (U), среднее значение функции конкурентного сходства (F_s) и меру Фишера (Q), которая пропорциональна 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.

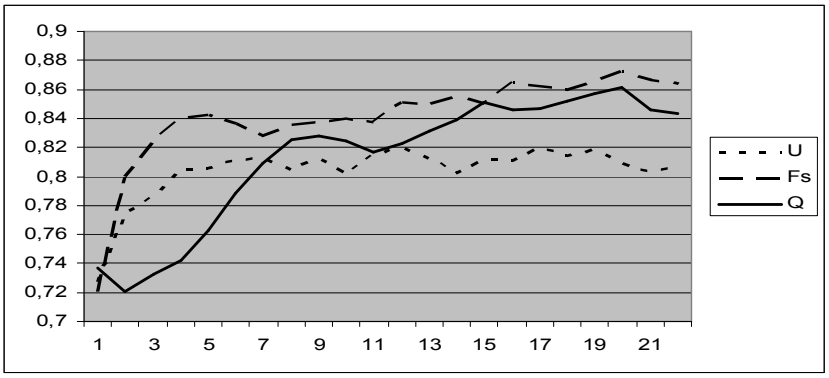

Рис. 2. Результаты выбора подсистем признаков при использовании трех критериев: по числу ошибок (U), по функции принадлежности (F_s) и по критерию Фишера Q.

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 F_s с результатами, полученными на обучении. Это свидетельствует о высоких прогностических свойствах критерия F_s.

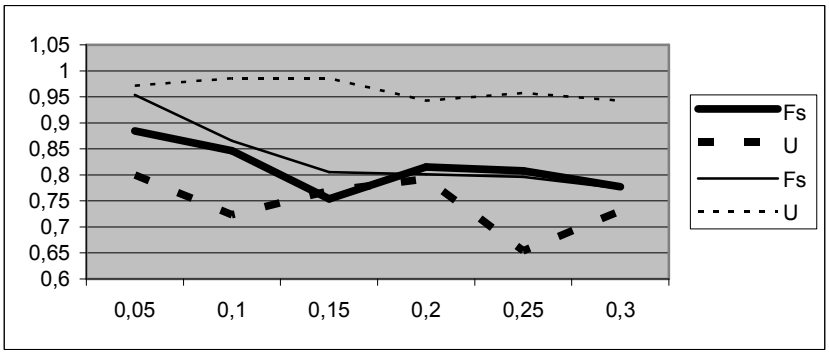

Рис. 3. Результаты обучения и распознавания по критериям U и F_s при разных уровнях шумов. Тонкие линии – обучение, жирные – контроль.

4 Оценка «пригодности» выбранных подсистем

Для оценки «пригодности» выбранных подсистем признаков мы проводили сравнение результатов, полученных при решении реальной задачи с результатами, полученными на таблице того же размера, но полученной из исходной таблицы путем случайной перестановки значений каждого признака. Такая перестановка разрушает имеющиеся зависимости между описывающими и целевым признаками. По этим чисто шумовым данным для каждой размерности подсистем n выбирались наиболее информативные признаки и определялись значения критерия F_s. Такие эксперименты повторялись много кратно и в результате можно было увидеть границы «случайного коридора» для получаемых значений F_s. Затем верхняя граница этого коридора сравнивалась со значением F_s для подсистем, найденных по исходной таблице. Если $F_s > F_s^{\max}$, то подсистемы признаков, найденные по исходной таблице могут считаться неслучайными. Если же 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.

5 Построение решающего правила

Для более быстрого и качественного распознавания для каждого из образов необходимо выбрать подмножество их объектов-эталонов (столпов), сходство с которыми будет использоваться для распознавания контрольных объектов. По существу столпы играют ту же роль, что и опорные вектора в известном методе SVM [3]. Один из методов выбора опорных векторов реализован в алгоритме STOLP [2]. В качестве опорных векторов эти алгоритмы выбирают объекты с наиболее высоким индивидуальным риском быть неправильно распознанными. Такие столпы обычно располагаются в области пересечения образов и защищают самих себя. Тот факт, что они могут использоваться в качестве эталонов при распознавании других объектов обучающей выборки, является, по существу, побочным. Сколько объектов защищает данный опорный вектор, и хорошо ли он их защищает, никак не влияет на процесс выбора опорных векторов. Но классика говорит, что если обучающая выборка образа подчиняется нормальному закону, то оптимальным эталоном является объект, находящийся не около разделяющих границ, а в точке математического ожидания образа.

Хотелось бы, чтобы алгоритм выбора столпов выдавал решения, адаптированные к ситуации. Если распределения унимодальны и нормальны, столпы должны располагаться в центрах тяжести образов. Если распределения полимодальны и образы линейно не разделимы, столпы должны стоять в центрах мод. С ростом сложности распределения число столпов k будет увеличиваться.

Алгоритм FRiS-Stolp, использующий в процессе работы функции конкурентного сходства, обладает именно такими свойствами. Он нацелен на выбор минимального числа столпов, которые защищают не только самих себя, но обеспечивают заданную надежность защиты всех остальных объектов обучающей выборки. Первыми выбираются столпы, защищающие максимально возможное количество объектов с заданной надежностью. По этой причине при нормальных распределениях в первую очередь будут выбраны столпы, расположенные в точках математического ожидания.

5.1 Алгоритм FRiS-Stolp

Пусть решается задача распознавания «первый образ против всех остальных».

1. Проверяется вариант, при котором первый случайно выбранный объект ai является единственным столпом образа S1, а все другие образы в качестве столпов имеют все свои объекты. Для всех объектов aj≠ai первого образа находится расстояние r1j до своего столпа ai и расстояние r2j до ближайшего объекта чужого образа. По этим расстояниям вычисляется значение FRiS-функции для каждого объекта aj первого образа. Находим те mi объектов первого образа, значение функций принадлежности F которых выше заданного порога F*, например, F*=0. По этим mi объектам вычисляем суммарное значение FRiS-функции Fi, которое характеризует пригодность объекта ai на роль столпа.
2. Аналогичную процедуру повторяем, назначая в качестве столпа все M объектов первого образа по очереди.
3. Находим объект ai с максимальным значением Fi и объявляем его первым столпом A11 первого кластера C11 первого образа S1.
4. Исключаем из первого образа mi объектов, входящих в первый кластер.
5. Для остальных объектов первого образа находим следующего столпа повторением пп 1- 4.
6. Процесс останавливается, если все объекты первого образа оказались включенными в свои кластеры .
7. Восстанавливаем все объекты образа S1 и для образа S2 выполняем пп 1-6.
8. Повторяем пп.1-7 для всех остальных образов.

На этом шаге заканчивается первый этап поиска столпов. Каждый столп Ai защищает подмножество объектов mi своего кластера Ci. Однако столпы были выбраны в условиях, когда им противостояли все объекты конкурирующих образов. Теперь образы представлены только своими столпами. Возможно, что в этих новых условиях некоторые (периферийные в своем кластере) объекты получили бы большее значение F, если бы у них была возможность

присоединиться не к первому, а к одному из последующих столпов. Предоставим объектам такую возможность. Для этого восстанавливаем все объекты обучающей выборки и распознаем их принадлежность к кластерам в условиях, когда функция принадлежности определяется по нормированным расстояниям до ближайшего своего и ближайшего чужого столпов.

Если состав некоторого кластера Ci изменился, то среди его объектов нужно повторить соревнование на лучшее исполнение роли столпа, как это делалось в п.1. Отличие будет состоять в том, что в соревновании участвуют только объекты этого кластера, и кандидат в столпы будет конкурировать со столпами чужих образов.

После завершения всех этих процедур вычисляется среднее значение функций конкурентного сходства *Fs* всех объектов обучающей выборки со своими столпами. Эта величина может служить мерой качества обучения.

Процесс распознавания контрольных объектов очень прост. Оцениваются расстояния r1 и r2 до двух ближайших столпов, принадлежащих разным образам, и выбирается тот образ i, значение Fi со столпом которого максимально. Тот же результат распознавания мы получили бы, если бы принимали решение в пользу образа, расстояние до столпа которого минимально. Но величина Fi оказалась полезной для того, чтобы получить ответ на простой и важный вопрос: «Если некоторый объект Z распознан в качестве представителя образа i, то какова достоверность этого конкретного решения?»

5.2 Оценка надежности распознавания конкретных объектов

Вопрос о том, есть ли связь между значением функции конкурентного сходства, вычисленной для конкретного объекта, и надежностью его распознавания (вероятностью неправильного распознавания), мы изучали на массиве реальных плохо обусловленных данных с неизвестными законами распределений. Эксперимент состоял в следующем.

1. Исходная выборка A, состоящая из 600 объектов в 1024-мерном пространстве, принадлежащих двум классам (по 300 объектов в классе) делилась на обучающую A_o (по 100 объектов на класс) и контрольную $A_к$ (по 200 объектов на класс). Методом Cross Validation на обучающей выборке оценивалась средняя величина ошибки распознавания Ps, а также строилась зависимость W между величиной FRiS-функции Fi и вероятностью ошибочного распознавания Pi объекта ai с таким значением FRiS-функции.
2. Для этого обучающая выборка случайным образом делилась на две равномощные подвыборки: A_{o1} и A_{o2}. На объектах из A_{o1} алгоритмом FRiS-Stolp строился набор столпов, по которым распознавались объекты из A_{o2}. Для каждого объекта ai из A_{o2} фиксировался результат распознавания («правильно»-«ошибка») и величина Fi для него.
3. Шаг 2. повторялся, пока не была набрана достаточная статистика для определения доли ошибочно распознанных объектов для небольших диапазонов изменения FRiS-функции. Полученная в результате гистограмма и была использована в качестве оценки W_o зависимости между величинами Fi и Pi объекта ai.
4. Затем на всей обучающей выборке A_o алгоритмом FRiS-Stolp строился набор столпов, по которым распознавалась вся контрольная выборка $A_к$. Для объектов контрольной выборки аналогичным способом строилась «реальная» зависимость W_k доли ошибок P при разных значения FRiS-функций F.

На рис. 4 приводится сравнение W_real с W_st. Здесь же приводится средняя вероятность ошибочного распознавания Ps, которая обычно используется при распознавании образов статистическими методами. Очевидно, что построенная на обучающей выборке зависимость между Fi и Pi с высокой надежностью прогнозирует вероятность ошибочного распознавания конкретных объектов контрольной выборки и эти оценки намного точнее, чем оценки «в среднем» по всей выборке.

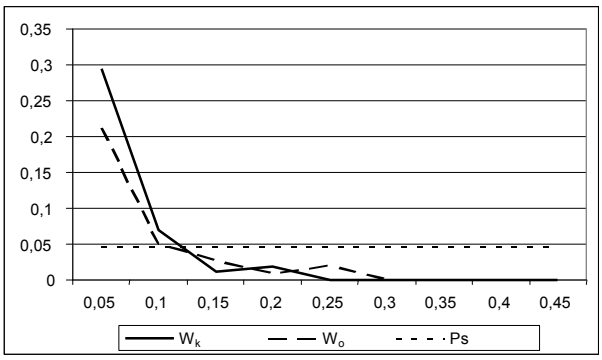

Рис.4. Сравнение зависимости W_o вероятности ошибочного распознавания от величины FRiS-функции, построенной на основе обучающей выборки, и аналогичной зависимости W_k на контрольной выборке.

6 Задача комбинированного типа DX

Выше рассмотрены две задачи основного типа – выбора информативных признаков (X) и построения решающих правил (D). Более интересна такая задача, в которой эти две задачи нужно решать одновременно: требуется найти такое подпространство признаков, в котором качество решающего правила было бы максимальным. Так формулируется задача комбинированного типа DX [2]. Применение функции конкурентного сходства позволяет построить простой алгоритм ее решения.

Общая схема алгоритма FRiS-DX такова:

1. С помощью того или иного алгоритма направленного перебора, например, с помощью алгоритма AdDel [2], выбирается очередной вариант признакового подпространства. В этом подпространстве алгоритм FRiS-Stolp проводит обучение, т.е. находит эталоны (столпы). Вычисляется среднее значение функций конкурентного сходства Fs всех объектов выборки со своими столпами. Величина Fs служит оценкой качества обученности системы.
2. Особенность алгоритма AdDel состоит в том, что в процессе увеличения размерности n выбираемых подпространств качество Fs обучения растет, затем рост прекращается и начинается его уменьшение. Точка перегиба функции Fs=f(n) указывает на наиболее предпочтительный вариант решения задачи. Рекомендуется использовать n выбранных в этой точке признаков. Они обеспечивают максимальное (или близкое к нему) качество обучения и, следовательно, и качество будущего распознавания контрольной выборки.

В многочисленных экспериментах на модельных и реальных данных сравнивались два варианта решения задачи обучения: в исходном пространстве признаков и в подпространстве, выбираемом алгоритмом FRiS-DX. Всегда результаты второго варианта оказывались существенно лучшими. При размерности исходного пространства в сотни и тысячи признаков размерность выбранного подпространства обычно не превышала нескольких единиц или первых десятков. При этом количество столпов также всегда уменьшалось, а среднее значение качества Fs заметно увеличивалось.

7 Заключение

Проведенные исследования позволяют сделать следующие выводы:

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

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

3. Использование функции конкурентного сходства позволяет на этапе распознавания сравнивать новые объекты не со всеми объектами обучающей выборки, а лишь с некоторыми эталонными образцами. Эти эталонные образцы строятся алгоритмом FRiS-Stolp и служат кратким описанием своих образов.

4. Значение функции F сходства контрольного объекта с эталоном своего образа дает возможность сопроводить результат распознавания оценкой правильности этого результата.

Литература

[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.: *Изд. ИМ СО РАН*, Новосибирск, 1999, 270 с.

[3] E. Boser, I.M. Guyon, and V.N. Vapnik A training algorithm for optimal margin classifiers.: *5th Annual ACM Workshop on COLT,* 1992, Pittsburgh, p. 144-152.

==========

Работа выполнена при поддержке РФФИ, грант №01-05-00241