УДК 004.93

С.А. СУББОТИН

КОМПЛЕКС ХАРАКТЕРИСТИК И КРИТЕРИЕВ СРАВНЕНИЯ ОБУЧАЮЩИХ ВЫБОРОК ДЛЯ РЕШЕНИЯ ЗАДАЧ ДИАГНОСТИКИ И РАСПОЗНАВАНИЯ ОБРАЗОВ

Abstract. *The actual problem of criteria set development for evaluating the training sample quality in the problems of diagnostics and pattern recognition on the features is solved in the paper. The experiments were conducted for the study of implementation of proposed criteria in practical problems solving. It confirms the practical usefulness of the developed criteria and software.*
Key words: *training set, training set quality, pattern recognition, diagnostics.*

Анотація. *Вирішено актуальне завдання розроблення комплексу критеріїв для оцінювання якості навчальних вибірок у задачах діагностики та розпізнавання образів за ознаками. Проведено експерименти з дослідження програмної реалізації запропонованих критеріїв при вирішенні практичних завдань, що підтвердили практичну корисність розробленого математичного забезпечення.*
Ключові слова: *навчальна вибірка, якість навчальної вибірки, розпізнавання образів, діагностика.*

Аннотация. *Решена актуальная задача разработки комплекса критериев для оценивания качества обучающих выборок в задачах диагностики и распознавания образов по признакам. Проведены эксперименты по исследованию программной реализации предложенных критериев при решении практических задач, подтвердившие практическую полезность разработанного математического обеспечения.*
Ключевые слова: *обучающая выборка, качество обучающей выборки, распознавание образов, диагностика*

1. Введение

Автоматизация процессов принятия решений в задачах диагностики и распознавания образов, как правило, предполагает необходимость решения задачи построения модели зависимости принимаемого решения от наблюдаемых переменных по прецедентам.

Для решения данной задачи применяют широкий арсенал методов математической статистики и вычислительного интеллекта, в частности, искусственные нейронные сети, нечеткие системы, деревья решений, методы распознавания образов, кластер-анализ [1–3].

Однако, несмотря на различия в обработке данных и структуре моделей, присущие разным методам, общим для них является использование обучающей выборки наблюдений для структурно-параметрической идентификации модели принятия решений.

При этом возникают две задачи:

– задача выбора метода, способного решить задачу наилучшим образом при наименьших затратах машинных и человеческих ресурсов;

– задача формирования такой выборки из имеющегося набора наблюдений, которая позволила бы синтезировать модель принятия решений наилучшим образом при наименьших затратах ресурсов.

Целью данной работы является создание комплекса критериев, характеризующих обучающую выборку с различных сторон и отражающих наиболее важные для моделирования свойства выборки. Это позволит обеспечить решение поставленных задач, а также существенным образом автоматизировать выбор метода построения модели из имеющихся в наборе и выбор экземпляров для включения в обучающую выборку.

2. Постановка задачи и анализ литературы

Пусть мы имеем обучающую выборку $< x, y >$, состоящую из экземпляров $x = \{x^s\},\ s = 1,2,...,S$, характеризующихся набором значений признаков $x^s = \{x_i^s\},\ i = 1,2,...,N$, которым сопоставлены

© Субботин С.А., 2010

ISSN 1028-9763. Математичні машини і системи, 2010, № 1

значения выходного признака $y = \{y^s\}$, где s – номер экземпляра выборки, N – количество описательных (входных) признаков, характеризующих экземпляры выборки, S – количество экземпляров в выборке. Для задач классификации обозначим число классов K.

Необходимо разработать комплекс критериев, отражающих наиболее важные свойства выборки для решения задач диагностики и распознавания образов.

Важнейшими свойствами выборки для решения задач распознавания образов являются [4–6]:

– репрезентативность – характеризует представительность выборки по отношению к генеральной совокупности (на практике данное свойство при неизвестных характеристиках генеральной совокупности обеспечивается достаточностью объема и полнотой выборки);

– полнота выборки определяется обеспеченностью классов экземплярами;

– размерность – характеризует, с одной стороны, пространственную сложность выборки, а с другой – минимальное количество операций обработки выборки;

– противоречивость – характеризует количество одинаковых объектов выборки, принадлежащих к разным классам;

– равномерность – показывает, насколько равномерно распределены экземпляры выборки по классам;

– компактность расположения классов в пространстве признаков – отражает простоту решения задачи распознавания (чем компактнее расположены экземпляры каждого класса, тем проще построить распознающую модель);

– сложность – характеризует затраты ресурсов памяти (пространственная сложность) и вычислительных ресурсов (вычислительная сложность) для обработки выборки.

Для некоторых из данных свойств ранее были предложены численные критерии, характеризующие их [4–7]. Однако известные критерии не отражают всей полноты свойств обучающих выборок, а также применимы не для всех задач (например, применимы только для задач с вещественной выходной переменной [6,7]).

Поэтому представляется целесообразным проанализировать и доработать известные критерии, а также разработать новые характеристики для формирования комплекса показателей, способного охарактеризовать важнейшие свойства выборки.

3. Критерии сравнения и характеристики обучающей выборки

Будем характеризовать свойства обучающей выборки $< x, y >$ с помощью следующего набора характеристик.

Размерность выборки определим как $Dm = NS$.

Данный показатель может изменяться от 1 до некоторой константы, поскольку число признаков и число экземпляров в обучающей выборке должны быть конечны. Тем не менее, для формирования обобщенного критерия данный критерий оказывается неудобным из-за плавающей верхней границы. Для устранения данного недостатка будем использовать относительную размерность выборки.

Относительную размерность выборки определим как $Dr = 1 - \exp(-\ln(Dm))$.

Величина Dr будет принимать значения в диапазоне $[0,1]$. При этом она будет чувствительной к малым размерностям, что практически очень полезно и удобно для сравнения различных выборок, в том числе для автоматизации процесса формирования выборки на основе интегрального критерия качества.

Некоторые из рассматриваемых далее критериев требуют задания выходной переменной как номера класса. Поэтому в задачах оценивания, где выходная переменная является вещественной, применение данных критериев предполагает выделение псевдоклассов, для чего можно использовать разбиение диапазона значений выходной переменной на равномерные интервалы:

$$y^s = round\left(1 + \frac{\left(y^s - \min\limits_{p=1,2,...,S}\left\{y^p\right\}\right)\left(round(\ln S) - 1\right)}{\max\limits_{p=1,2,...,S}\left\{y^p\right\} - \min\limits_{p=1,2,...,S}\left\{y^p\right\}}\right),$$

где $round(a)$ – функция округления.

Косвенно полноту и равномерность выборки предлагается характеризовать такими показателями, как

– оценка априорной вероятности (частоты) q-го класса по выборке:

$$P(y = q) = \frac{S^q}{S},\ q = 1, 2, ..., K,$$

где S^q – количество экземпляров выборки, принадлежащих q-му классу, K – количество классов, выделяемое в данной задаче;

– минимальная частота класса в выборке:

$$P_{\min} = S^{-1} \min_{q=1,2,...,K}\left\{S^q\right\};$$

– среднее отклонение частоты класса по выборке:

$$\sigma = \sum_{q=1}^{K}\left(\frac{1}{K} - \frac{S^q}{S}\right)^2.$$

Данная величина будет изменяться в диапазоне от нуля (если классы имеют одинаковые частоты) до некоторой положительной константы (если классы имеют неодинаковые частоты). Причем она будет тем больше, чем выше неравномерность частот классов;

– инверсное нормированное среднее отклонение частоты класса по выборке:

$$\sigma_{норм.} = \exp\left(-\sum_{q=1}^{K}\left(\frac{1}{K} - \frac{S^q}{S}\right)^2\right).$$

Данная величина будет изменяться в диапазоне от нуля (если классы имеют неодинаковые частоты) до единицы (если классы имеют одинаковые частоты). Причем она будет тем меньше, чем выше неравномерность частот классов.

Для оценки неравномерности обучающей выборки в [4] используется показатель

$$Rg = \sqrt{\sum_{q=1}^{K}\left(S^q - \frac{1}{K}\sum_{k=1}^{K} S^k \right)^2}.$$

Его недостатком является то, что данный показатель имеет подвижную верхнюю границу в области значений. Выполнив нормирование, получим относительную характеристику неравномерности обучающей выборки:

$$Rg' = \frac{1}{S}\sqrt{\sum_{q=1}^{K}\left(S^q - \frac{1}{K}\sum_{k=1}^{K} S^k \right)^2}.$$

Полученный показатель будет принимать значения в диапазоне от 0 до 1: чем меньше будет его значение, тем более равномерным будет распределение экземпляров выборки по классам.

Соответственно определим характеристику относительной равномерности обучающей выборки как $Nr = 1 - Rg'$.

Полученный показатель будет принимать значения в диапазоне от 0 до 1 : чем больше будет его значение, тем более равномерным будет распределение экземпляров выборки по классам.

Равномерность распределения экземпляров выборки по оси значений i -го признака определим как

$$Ev_i = \frac{1}{S}\sum_{g=1}^{S}\omega_{ig}, \quad \omega_{ig} = \begin{cases} \left(\sum_{s=1}^{S}\omega_i\left(x^s, g\right)\right)^{-1}, \sum_{s=1}^{S}\omega_i\left(x^s, g\right) > 0, \\ 0, \sum_{s=1}^{S}\omega_i\left(x^s, g\right) = 0, \end{cases}$$

где $\omega_i(x^s, g) = \begin{cases} 1, (g-1) \le \dfrac{(x_i^s - \min\limits_{p=1,2,...,S}(x_i^p))S}{\max\limits_{p=1,2,...,S}(x_i^p) - \min\limits_{p=1,2,...,S}(x_i^p)} \le g, \\ 0, \textit{в противном случае} \end{cases}$

либо

$$\omega_i(x^s, g) = \begin{cases} \exp\left(-\left(x_i^s - \frac{1}{2S}(2g-1)\left(\max\limits_{p=1,2,...,S}(x_i^p) - \min\limits_{p=1,2,...,S}(x_i^p) \right) \right)^2 \right), (g-1) \le \dfrac{(x_i^s - \min\limits_{s=1,2,...,S}(x_i^s))S}{\max\limits_{s=1,2,...,S}(x_i^s) - \min\limits_{s=1,2,...,S}(x_i^s)} \le g, \\ 0, \textit{в противном случае}. \end{cases}$$

Чем ближе значение Ev_i к единице, тем равномернее распределены экземпляры по оси значений i -го признака. В свою очередь, чем ближе значение Ev_i к нулю, тем менее равномерно распределены экземпляры по оси значений i -го признака.

Неравномерность распределения экземпляров выборки по оси значений i -го признака:

$$NEv_i = 1 - Ev_i.$$

Чем ближе значение NEv_i к единице, тем менее равномерно распределены экземпляры по оси значений i-го признака. В свою очередь, чем ближе значение NEv_i к нулю, тем равномернее распределены экземпляры по оси значений i-го признака.

Равномерность покрытия экземплярами выборки признакового пространства определим как

$$Ev = \frac{1}{N}\sum_{i=1}^{N} Ev_i.$$

Чем ближе значение Ev к единице, тем равномернее распределены экземпляры в пространстве признаков, что лучше с точки зрения адекватности отображения свойств генеральной совокупности выборки в рассматриваемой части признакового пространства, однако хуже с точки зрения возможной избыточности выборки. В свою очередь, чем ближе значение Ev к нулю, тем менее равномерно распределены экземпляры в пространстве признаков, что хуже с точки зрения адекватности отображения свойств генеральной совокупности выборки в рассматриваемой части признакового пространства.

Неравномерность покрытия экземплярами выборки признакового пространства:

$$NEv = 1 - Ev.$$

Чем ближе значение NEv к единице, тем менее равномерно распределены экземпляры по оси значений i-го признака. В свою очередь, чем ближе значение NEv к нулю, тем равномернее распределены экземпляры по оси значений i-го признака.

Равномерность распределения экземпляров q-го класса по оси значений i-го признака:

$$Ev_i^q = \frac{1}{S^q}\sum_{g=1}^{S^q} \omega_{ig}^q, \;\; \omega_{ig}^q = \begin{cases} \left(\sum_{s=1}^{S}\left\{\omega_i^q\left(x^s, g\right) \mid y^s = q\right\}\right)^{-1}, \sum_{s=1}^{S}\left\{\omega_i^q\left(x^s, g\right) \mid y^s = q\right\} > 0, \\ 0, \sum_{s=1}^{S}\left\{\omega_i^q\left(x^s, g\right) \mid y^s = q\right\} = 0, \end{cases}$$

где $\omega_i^q(x^s, g) = \begin{cases} 1, y^s = q, (g-1) \le \dfrac{(x_i^s - \min\limits_{p=1,2,...,S}(x_i^p))S}{\max\limits_{p=1,2,...,S}(x_i^p) - \min\limits_{p=1,2,...,S}(x_i^p)} \le g, \\ 0, \textit{в противном случае} \end{cases}$

либо

$$\omega_i^q(x^s, g) = \begin{cases} \exp\left(-\left(x_i^s - \dfrac{1}{2S}(2g-1)\left(\max\limits_{p=1,2,...,S}(x_i^p) - \min\limits_{p=1,2,...,S}(x_i^p)\right)\right)^2\right), \\ \qquad y^s = q, (g-1) \le \dfrac{(x_i^s - \min\limits_{s=1,2,...,S}(x_i^s))S}{\max\limits_{s=1,2,...,S}(x_i^s) - \min\limits_{s=1,2,...,S}(x_i^s)} \le g. \\ 0, \textit{в противном случае.} \end{cases}$$

Чем ближе значение Ev_i^q к единице, тем равномернее распределены экземпляры по оси значений i-го признака, и, следовательно, ситуация хуже с точки зрения гипотезы о компактности классов и разделяющих свойств i-го признака. В свою очередь, чем ближе значение Ev_i^q к нулю, тем менее равномерно распределены экземпляры по оси значений i-го признака, и, следовательно, ситуация лучше с точки зрения гипотезы о компактности классов и разделяющих свойств i-го признака.

Неравномерность распределения экземпляров q-го класса по оси значений i-го признака:

$$NEv_i^q = 1 - Ev_i^q .$$

Чем ближе значение NEv_i^q к единице, тем менее равномерно распределены экземпляры q-го класса по оси значений i-го признака. В свою очередь, чем ближе значение NEv_i^q к нулю, тем равномернее распределены экземпляры q-го класса по оси значений i-го признака.

Равномерность покрытия экземплярами q-го класса признакового пространства:

$$Ev^q = \frac{1}{N}\sum_{i=1}^{N} Ev_i^q .$$

Чем ближе значение Ev^q к единице, тем равномернее распределены экземпляры q-го класса в пространстве признаков, что хуже с точки зрения гипотезы о компактности классов. В свою очередь, чем ближе значение Ev^q к нулю, тем менее равномерно распределены экземпляры в пространстве признаков, что лучше с точки зрения гипотезы о компактности классов.

Неравномерность покрытия экземплярами q-го класса признакового пространства:

$$NEv^q = 1 - Ev^q .$$

Чем ближе значение NEv^q к единице, тем менее равномерно распределены экземпляры q-го класса в пространстве признаков.

Средняя равномерность покрытия экземплярами классов признакового пространства будет определяться как

$$\overline{E}v = \frac{1}{K}\sum_{q=1}^{K} Ev^q .$$

Минимальный уровень равномерности покрытия экземплярами классов признакового пространства будет определяться как

$$\breve{E}v = \min_{q=1,2,..,K}\left\{Ev^q\right\}.$$

Повторяемость обучающей выборки, согласно [4], может быть определена как показатель, характеризующий количество одинаковых экземпляров, принадлежащих к одному и тому же классу. Пронормировав, формально это можно представить как

$$Rp(x,y) = \frac{2}{S(S-1)}\sum_{s=1}^{S}\sum_{g=s+1}^{S}\tau(x,y,s,g),$$

где $\tau(x, y, s, g) = \begin{cases} 1, y^s = y^g, \forall i = 1,2,..., N : x_i^s = x_i^g, \\ 0, \textit{в противном случае}. \end{cases}$

Величина Rp будет минимальной (равной нулю) в случае, если все экземпляры обучающей выборки отличны друг от друга, и максимальной (равной единице), если все экземпляры одинаковы.

Однако такой показатель будет реагировать только на абсолютные совпадения обучающих примеров. На практике же часто приходится иметь дело с выборками, в которых содержатся не одинаковые, но близкие по свойствам (почти одинаковые) экземпляры одного класса. Для учета подобных случаев переопределим показатель τ как

$$\tau(x, y, s, g) = \begin{cases} \exp\left(-\alpha \sum_{i=1}^{N} (x_i^s - x_i^g)^2 \right), y^s = y^g, \\ 0, y^s \neq y^g. \end{cases}$$

Здесь α – коэффициент, регулирующий положение границы локальной близости экземпляров одного класса, $\alpha > 0$. В простейшем случае можно положить $\alpha = 1$.

Полученная формула будет применима для задач распознавания образов, однако будет мало пригодна для задач оценивания. Для задач, где выходная переменная принимает вещественные значения в некотором диапазоне, переопределим показатель τ как

$$\tau(x, y, s, g) = \begin{cases} \exp\left(-\alpha \sum_{i=1}^{N} (x_i^s - x_i^g)^2 \right), \left| y^s - y^g \right| \leq \delta, \\ 0, \left| y^s - y^g \right| > \delta. \end{cases}$$

где δ – константа, регулирующая чувствительность для определения подобия значений выходной переменной, $\delta > 0$. Значение константы δ предлагается автоматически определять предварительно на основе формулы

$$\delta = \frac{1}{2S - 1} \left(\sqrt{\left(\max_{s=1,2,...,S} \{y^s\} - \min_{s=1,2,...,S} \{y^s\} \right)^2} + \sqrt{\sum_{s=1}^{S} \sum_{g=s+1}^{S} (y^s - y^g)^2} \right).$$

В качестве противоположной характеристики выборки по отношению к повторяемости определим уникальность экземпляров выборки как $Rn = 1 - Rp$.

При построении распознающих моделей часто выдвигается требование независимости входных переменных. Для оценивания качества выборки с точки зрения данного требования будем использовать показатели:

– усредненной независимости входных переменных: $\bar{I}dp = 1 - \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left| r_{i,j} \right|$,

где $r_{i,j}$ – коэффициент парной корреляции, для вещественных признаков определяемый по формуле

$$r_{i,j}=\frac{\sum_{s=1}^{S}\left(x_i^s-\sum_{g=1}^{S}x_i^g\right)\left(x_j^s-\sum_{g=1}^{S}x_j^g\right)}{\sqrt{\sum_{s=1}^{S}\left(x_i^s-\sum_{g=1}^{S}x_i^g\right)^2\left(x_j^s-\sum_{g=1}^{S}x_j^g\right)^2}};$$

– минимальной независимости входных переменных: $\breve{I}dp=1-\max\limits_{\substack{i=1,2,...,N;\\ j=i+1,...,N}}\left|r_{i,j}\right|$;

– максимальной независимости входных переменных: $\hat{I}dp=1-\min\limits_{\substack{i=1,2,...,N;\\ j=i+1,...,N}}\left|r_{i,j}\right|$.

Наряду с независимостью входных переменных между собой, при решении задач построения моделей выдвигается требование наличия связи между выходной и входными переменными, причем предпочтительнее линейная связь. Для характеристики отображения в выборке связи входных и выходной переменных предлагается использовать показатели:

– максимальной линейной связи входных и выходной переменных: $\hat{Y}dp=\max\limits_{i=1,2,...,N}\left|r_{i,y}\right|$,

где $r_{i,y}$ – коэффициент парной корреляции i-го признака и выходного признака;

– средней линейной связи входных и выходной переменных: $\bar{Y}dp=\frac{1}{N}\sum_{i=1}^{N}\left|r_{i,y}\right|$;

– комбинированные показатели независимости входных переменных и линейности связи с выходной переменной:

$$\hat{I}_Y=\max_{\substack{i=1,2,...,N;\\ j=i+1,...,N}}\left\{\left|r_{i,y}\right|\left(1-\left|r_{i,j}\right|\right)\right\},\ \breve{I}_Y=\min_{\substack{i=1,2,...,N;\\ j=i+1,...,N}}\left\{\left|r_{i,y}\right|\left(1-\left|r_{i,j}\right|\right)\right\},\ \bar{I}_Y=\frac{2}{N(N-1)}\sum_{i=1}^{N}\left(\left|r_{i,y}\right|\sum_{j=i+1}^{N}\left(1-\left|r_{i,j}\right|\right)\right).$$

Компактность расположения экземпляров q-го класса по i-му признаку:

$$Co_i^q=1-\frac{2\sum_{s=1}^{S}\sum_{g=s+1}^{S}\left\{(x_i^s-x_i^g)^2\mid y^s=y^g=q\right\}}{S^q(S^q-1)\left(\max\limits_{s=1,2,..,S}\{x_i^s\mid y^s=q\}-\min\limits_{s=1,2,..,S}\{x_i^s\mid y^s=q\}\right)^2}.$$

Чем больше значение Co_i^q, тем, в среднем, компактнее расположены экземпляры q-го класса по i-му признаку.

Компактность расположения экземпляров q-го класса:

$$Co^q=1-\frac{2\sum_{s=1}^{S}\sum_{g=s+1}^{S}\sum_{i=1}^{N}\left\{(x_i^s-x_i^g)^2\mid y^s=y^g=q\right\}}{S^q(S^q-1)\sum_{i=1}^{N}\left(\max\limits_{s=1,2,..,S}\{x_i^s\mid y^s=q\}-\min\limits_{s=1,2,..,S}\{x_i^s\mid y^s=q\}\right)^2}.$$

Чем больше значение Co^q, тем, в среднем, компактнее расположены экземпляры *q*-го класса в пространстве признаков.

Компактность расположения экземпляров q-го и p-го классов по i-му признаку:

$$Co_i^{q,p} = 1 - \frac{2\sum_{s=1}^{S}\sum_{g=s+1}^{S}\left\{(x_i^s - x_i^g)^2 \mid (y^s = q \vee y^s = p) \vee (y^g = q \vee y^g = p)\right\}}{(S^p + S^q)(S^p + S^q - 1)\left(\max\limits_{s=1,2,..,S}\{x_i^s \mid (y^s = q \vee y^s = p)\} - \min\limits_{s=1,2,..,S}\{x_i^s \mid (y^s = q \vee y^s = p)\}\right)^2}.$$

Чем больше значение $Co_i^{q,p}$, тем, в среднем, сложнее отделить q-й и p-й классы друг от друга по i-му признаку, но легче отделить в совокупности q-й и p-й классы от остальных классов по i-му признаку.

Компактность расположения экземпляров q-го и p-го классов в пространстве признаков:

$$Co(q,p) = 1 - \frac{2\sum_{s=1}^{S}\sum_{g=s+1}^{S}\sum_{i=1}^{N}\left\{(x_i^s - x_i^g)^2 \mid (y^s = q \vee y^s = p) \vee (y^g = q \vee y^g = p)\right\}}{(S^p + S^q)(S^p + S^q - 1)\sum_{i=1}^{N}\left(\max\limits_{s=1,2,..,S}\{x_i^s \mid (y^s = q \vee y^s = p)\} - \min\limits_{s=1,2,..,S}\{x_i^s \mid (y^s = q \vee y^s = p)\}\right)^2}.$$

Чем больше значение $Co(q,p)$, тем, в среднем, сложнее отделить q-й и p-й классы друг от друга, но легче отделить в совокупности q-й и p-й классы от остальных классов.

Усредненная компактность классов: $\overline{Co} = \frac{1}{K}\sum_{q=1}^{K} Co^q.$

Чем больше значение усредненной компактности классов, тем теснее внутри каждого класса расположены экземпляры, что свидетельствует в пользу гипотезы о компактности классов.

Минимальная компактность классов: $Co^{\min} = \min\limits_{q=1,2,..,K}(Co^q).$

Чем больше значение минимальной компактности классов, тем теснее внутри каждого класса расположены экземпляры, что свидетельствует в пользу гипотезы о компактности классов.

Отделимость q-го класса:

$$Se^q = \frac{1}{1 + \min\limits_{p=1,2,...,K} Co(q,p)}.$$

Чем меньше минимальная совместная компактность q-го класса со всеми остальными классами, тем более легко отделить экземпляры q-го от остальных классов. Следовательно, будет больше значение отделимости q-го класса.

Отделимость классов:

$$Se = \frac{1}{1 + \min\limits_{\substack{q=1,2,...,K;\\ p=q+1,2,...,K;}} Co(q,p)}.$$

Чем больше значение отделимости классов, тем более компактно расположен каждый из классов и сильнее его отделимость от других классов, что обуславливает применение методов распознавания, основанных на гипотезе о компактности.

Упрощенный показатель компактности-отделимости классов определим по формуле

$$SC = 1 - \exp\left(- \frac{\min\limits_{\substack{s \neq p,\\ s=1,2,...,S;\\ p=s+1,...,S}} \left\{ \sum\limits_{i=1}^{N} \left\{ \left(x_i^s - x_i^p \right)^2 \middle| y^s \neq y^p \right\} \right\}}{1 + \min\limits_{\substack{s \neq p,\\ s=1,2,...,S;\\ p=s+1,...,S}} \left\{ \sum\limits_{i=1}^{N} \left(x_i^s - x_i^p \right)^2 \right\}} \right).$$

Значения данного показателя будут расположены в интервале от 0 до 1: чем меньше значение критерия, тем более тесно расположены (более сложно разделимы) разные классы и тем менее сконцентрированы экземпляры одного и того же класса.

В [4] предложено характеризовать противоречивость обучающей выборки как (формула приведена в уточненном виде с подстановками)

$$Cnd = \frac{2}{S(S-1)} \sum_{s=1}^{S} \sum_{p=s+1}^{S} \frac{\sqrt{\sum\limits_{i=1}^{N} \left(C_i^{y^s} - C_i^{y^p} \right)^2}}{\sqrt{\sum\limits_{i=1}^{N} \left(C_i^{y^s} - C_i^{y^p} \right)^2} + \sqrt{\sum\limits_{i=1}^{N} \frac{\left(x_i^s - x_i^p \right)^2}{\frac{1}{S} \sum\limits_{g=1}^{S} \left(x_i^s - \bar{x}_i \right)^2}}},$$

где $\bar{x}_i = \frac{1}{S} \sum\limits_{s=1}^{S} x_i^s$, $C_i^q = \frac{1}{S^q} \sum\limits_{s=1}^{S} \left\{ x_i^s \mid y^s = q \right\}$.

Достоинством данного критерия является то, что его значения находятся в интервале от 0 до 1: чем больше значение критерия, тем более противоречивой является выборка. Недостатком критерия является его зависимость от гипотезы компактности образов: на практике образы могут быть представлены множеством кластеров, а также содержать взаимопроникновения. Этот критерий также не применим для задач с вещественным выходом.

Относительную противоречивость обучающей выборки будем оценивать по формуле

$$Ic = \frac{1}{S(S-1)} \sum_{s=1}^{S} \sum_{g=s+1}^{S} \tau'(x, y, s, g),$$

где $\tau'(x, y, s, g) = \begin{cases} 1, y^s \neq y^g, \forall i = 1,2,..., N : x_i^s = x_i^g, \\ 0, \textit{в противном случае} \end{cases}$ либо

$$\tau'(x, y, s, g) = \begin{cases} \exp\left(-\alpha \sum\limits_{i=1}^{N} (x_i^s - x_i^g)^2 \right), y^s \neq y^g, \\ 0, y^s = y^g \end{cases}$$ либо $$\tau'(x, y, s, g) = \begin{cases} \exp\left(-\alpha \sum\limits_{i=1}^{N} (x_i^s - x_i^g)^2 \right), \left| y^s - y^g \right| > \delta, \\ 0, \left| y^s - y^g \right| \leq \delta. \end{cases}$$

Показатель относительной противоречивости будет принимать значения в диапазоне от 0 до 1: чем меньше будет его значение, тем меньше доля одинаковых экземпляров, принадлежащих к разным классам.

В свою очередь, относительную непротиворечивость обучающей выборки определим как

$$Cn = 1 - Ic.$$

Показатель относительной непротиворечивости будет принимать значения в диапазоне от 0 до 1: чем больше будет его значение, тем меньше доля одинаковых экземпляров, принадлежащих к разным классам.

Сложность обучающей выборки $< x, y >$ для аппроксимации функции $y = f(x)$ в случае, когда выходная переменная является вещественной, может быть оценена с помощью константы Липшица [6, 7]:

$$L(x, y) = \max_{\substack{s=1,2,...,S;\\ g=s+1,...,S}} \left\{ \frac{\sqrt{(y^s - y^g)^2}}{\sqrt{\sum_{i=1}^{N} (x_i^s - x_i^g)^2}} \right\}.$$

Для задач распознавания, когда выходная переменная принимает дискретные значения, константа Липшица будет зависеть в основном от знаменателя. При этом следует учесть тот факт, что номера классов в числителе могут не выражать степень их различия. Поэтому определим сложность аппроксимации, модифицировав константу Липшица следующим образом:

$$L'(x, y) = \max_{\substack{s \neq g,\\ s=1,2,...,S;\\ g=s+1,...,S}} \left\{ \frac{1}{\sqrt{\sum_{i=1}^{N} (x_i^s - x_i^g)^2}} \right\} = \frac{1}{\sqrt{\min\limits_{\substack{s \neq g,\\ s=1,2,...,S;\\ g=s+1,...,S}} \left\{ \sum_{i=1}^{N} (x_i^s - x_i^g)^2 \right\}}}.$$

Рассмотренные показатели сложности выборки сильно зависят от размерностей входных и выходной переменной и не удобны в использовании при сравнении разных задач. Для устранения данного недостатка, а также оптимизации вычислений предлагается использовать модифицированные показатели сложности обучающей выборки (здесь также обеспечивается неравенство знаменателя нулю):

– для задач с вещественной выходной переменной:

$$L''(x, y) = \sqrt{\max_{\substack{s=1,2,...,S;\\ g=s+1,...,S}} \left\{ \frac{\nu_y (y^s - y^g)^2}{1 + \sum_{i=1}^{N} \nu_i (x_i^s - x_i^g)^2} \right\}},$$

где $\nu_i = \left(\frac{1}{\max\limits_{s=1,2,...,S} \{x_i^s\} - \min\limits_{s=1,2,...,S} \{x_i^s\}} \right)^2$, $\nu_y = \left(\frac{1}{\max\limits_{s=1,2,...,S} \{y^s\} - \min\limits_{s=1,2,...,S} \{y^s\}} \right)^2$;

– для задач с дискретной выходной переменной:

$$L''(x, y) = \sqrt{\frac{1}{1 + \min\limits_{\substack{s \neq g,\\ s=1,2,...,S;\\ g=s+1,...,S}} \left\{ \sum_{i=1}^{N} \nu_i (x_i^s - x_i^g)^2 \right\}}}.$$

Модифицированный показатель будет характеризовать относительную сложность аппроксимации зависимости по обучающей выборке. При этом его значения будут находиться в диапазоне от нуля до единицы: чем меньше будет значение показателя относительной сложности, тем лучше выборка будет подходить для решения задачи.

Для показателя относительной сложности определим альтернативный ему показатель относительной простоты аппроксимации зависимости по обучающей выборке как $Si = 1 - L''$.

Значения показателя относительной простоты аппроксимации зависимости по обучающей выборке будут находиться в диапазоне от нуля до единицы: чем больше будет значение показателя относительной простоты, тем лучше будет подходить выборка для решения задачи.

На основе комплекса рассмотренных характеристик возможно определить интегральные показатели качества обучающей выборки:

– критерий отбора экземпляров:

$$I_{\mathrm{Q}}^{\text{экз.}} = \frac{S_{\max}\sigma_{\text{норм.}}(Nr + Ev)}{2S(1+Rp)} \to \max,$$

где $S_{\max}$ – максимально возможное число экземпляров выборки;

– критерий отбора признаков:

$$I_{\mathrm{Q}}^{\text{призн.}} = \frac{N_{\max}}{18N} \frac{(\bar{I}dp + \hat{I}dp + \bar{Y}dp + \hat{Y}dp + \bar{I}y + \hat{I}y)(Se + Sc + \overline{C}o)}{1 + \frac{1}{6}(Ev + \overline{E}v + \breve{E}v)(Cnd + Ic)L''} \to \max,$$

где $N_{\max}$ – максимально возможное число признаков в выборке;

– обобщенный показатель качества выборки:

$$I_{\mathrm{Q}} = \frac{\sigma_{\text{норм.}} NrRn(\bar{I}dp + \hat{I}dp + \bar{Y}dp + \hat{Y}dp + \bar{I}y + \hat{I}y)(Se + Sc + \overline{C}o)}{18 + 3Dr(Ev + \overline{E}v + \breve{E}v)(Cnd + Ic)L''} \to \max.$$

4. Эксперименты и результаты

Разработанный комплекс критериев был программно реализован в виде библиотеки функций на языке пакета Matlab, которая использовалась для исследования практической применимости разработанных критериев.

Для исследования предложенного комплекса критериев и программного обеспечения, реализующего их, использовались выборки данных для задач: определения вида ирисов [8], автоматической классификации сельскохозяйственных растений на культурные и сорные по данным дистанционного зондирования [9], неразрушающей диагностики лопаток газотурбинных авиадвигателей [1], прогнозирования суммарного показателя качества жизни (СПКЖ) больных хроническим обструктивным бронхитом [10].

Характеристики выборок и расчетные значения критериев представлены в табл. 1.

Таблица 1. Характеристики и критерии сравнения обучающих выборок

Критерий	**Задача**					
	Ирисы Фишера	Распознавание растений		Диагностика лопаток		СПКЖ-бронхит
S	150	248	248	32	32	86
N	4	55	5	100	10	106
K	3	2	2	2	2	0
Dm	600	13640	1240	3200	320	9116
Dr	0,99833	0,99993	0,99919	0,99969	0,99687	0,99989
$P_{\min}$	0,33333	0,41935	0,41935	0,5	0,5	0,10465
σ	0	0,013007	0,013007	0	0	0,034749
$\sigma_{норм}$	1	0,98708	0,98708	1	1	0,96585
Rg	0	0	0	0	0	0
Rg'	0	0	0	0	0	0
Nr	1	1	1	1	1	1
Ev	0,085526	0,21092	0,24596	0,30663	0,36599	0,087689
NEv	0,91447	0,78908	0,75404	0,69337	0,63401	0,91231
$\overline{Ev}$	0,077894	0,34922	0,39197	0,28915	0,33444	0,084681
$\breve{E}v$	0,03125	0,32343	0,35024	0,23249	0,31756	0,060485
Rp	0,000089	0	0	0,0020161	0,0020161	0
Rn	0,99991	1	1	0,99798	0,99798	1
$\bar{I}dp$	0,40588	0,15256	0,12047	0,73728	0,35286	0,80346
$\breve{I}dp$	0,037135	0,000012	0,019812	0,089	0,0092664	0,02352
$\hat{I}dp$	0,88243	0,49649	0,21094	0,99996	0,86321	1
$\hat{Y}dp$	0,95655	0,16513	0,070793	0,65551	0,67929	0,93089
$\overline{Y}dp$	0,7787	0,05694	0,030351	0,21573	0,57528	0,38172
$\hat{I}_Y$	0,69056	0,05725	0,0069386	0,65165	0,5471	0,71621
$\breve{I}_Y$	0,035242	$0{,}31\cdot10^{-8}$	0,0002240	0,00016274	0,0057847	0
$\bar{I}_Y$	0,24716	0,011098	0,001956	0,18522	0,2197	0,23898
$\overline{Co}$	0,90486	0,91523	0,91265	0,88221	0,85429	0,86195
$Co^{\min}$	0,89057	0,89445	0,88996	0,8794	0,83283	0,81453
Se	0,56051	0,51382	0,51474	0,51786	0,53182	0,53194
SC	0,048771	0,000464	0,0024585	0,98999	0,63083	0,99981
Cnd	0,34042	0,010598	0,084844	0,11061	0,33531	–
Ic	0,006711	0	0	0	0	0
Cn	0,99329	1	1	1	1	1
L	–	–	–	–	–	0,0059119
L''	0,51803	0,1553	0,43095	0,16441	0,35889	0,15809
Si	0,48197	0,8447	0,56905	0,83559	0,64111	0,84191
$I_{\mathrm{Q}}^{\text{экз.}}$	0,54271	0,59764	0,61493	0,652	0,68162	0,52527
$I_{\mathrm{Q}}^{\text{призн.}}$	0,33129	0,074592	0,38343	0,45633	3,555	0,53811
I_Q	0,33126	0,073629	0,034407	0,45541	0,35481	0,51973

Как видно из табл. 1, разработанный комплекс критериев позволяет оценить качество обучающей выборки с различных сторон и на практике автоматизировать процесс формирования (выбора) обучающего множества для решения задач диагностики и распознавания образов.

В частности, предложенные обобщенные показатели качества обучающих выборок достаточно хорошо показывают уменьшение размерности задачи, повышение информативности ее описания, изменение сложности решения задачи и др.

5. Заключение

С целью автоматизации формирования обучающих множеств, отбора информативных признаков и выбора метода обучения распознавания образов в работе решена актуальная задача разработки математического обеспечения для оценивания свойств обучающих выборок.

Научная новизна работы заключается в том, что: получили дальнейшее развитие метод и критерий оценки сложности обучающей выборки на основе константы Липшица, модифицированный путем нормирования и учета специфики задачи, что позволяет оценивать сложность аппроксимации функции по выборке как для задач с вещественным, так и для задач с дискретным выходом; модифицирован показатель повторяемости обучающей выборки; впервые предложены интегральные показатели качества выборки (критерий отбора экземпляров, критерий отбора признаков, обобщенный показатель качества выборки), методы и критерии оценки относительной размерности выборки, относительной простоты аппроксимации зависимости, относительной противоречивости и непротиворечивости обучающей выборки, критерии отделимости классов, критерии компактности классов, упрощенный показатель компактности-отделимости классов, показатели независимости входных переменных и их связи с выходной переменной, характеристики равномерности и неравномерности выборки. Разработанные и модифицированные методы и критерии позволяют количественно выразить пригодность выборки для решения задач диагностики и распознавания образов.

Практическая ценность работы состоит в том, что разработано программное обеспечение, позволяющее для заданной обучающей выборки автоматически рассчитывать показатели ее качества, что дает возможность автоматически формировать обучающие множества и сравнивать их, а также выбирать наиболее эффективные для решения задачи методы.

Работа выполнена в рамках госбюджетной темы "Информационные технологии автоматизации распознавания образов и принятия решений для диагностики в условиях неопределенности на основе гибридных нечеткологических, нейросетевых и мультиагентных методов вычислительного интеллекта" кафедры программных средств Запорожского национального технического университета.

СПИСОК ЛИТЕРАТУРЫ

1. Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей / В.И. Дубровин, С.А. Субботин, А.В. Богуслаев, В.К. Яценко. – Запорожье: ОАО "Мотор–Сич", 2003. – 279 с.
2. Субботін С.О. Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей: Монографія / С.О. Субботін, А.О. Олійник, О.О. Олійник; під заг. ред. С.О. Субботіна. – Запоріжжя: ЗНТУ, 2009. – 375 с.
3. Субботін С.О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень: Навчальний посібник / Субботін С.О. – Запоріжжя: ЗНТУ, 2008. – 341 с.
4. Олешко Д.Н. Построение качественной обучающей выборки для прогнозирующих нейросетевых моделей / Д.Н. Олешко, В.А. Крисилов, А.А. Блажко // Штучний інтелект. – 2004. – № 3. – С. 567 – 573.

5. Крисилов В.А. Представление исходных данных в задачах нейросетевого прогнозирования / В.А. Крисилов, К.В. Чумичкин, А.В. Кондратюк. – Нейроинформатика–2003. – М.: МИФИ, 2003. – Ч. 1. – С. 184 –191.
6. Царегородцев В.Г. Оптимизация предобработки данных: константа Липшица обучающей выборки и свойства обученных нейронных сетей / В.Г. Царегородцев // Нейрокомпьютеры: разработка, применение. – 2003. – №7. – С. 3 – 8.
7. Царегородцев В.Г. Предобработка обучающей выборки, выборочная константа Липшица и свойства обученных нейронных сетей / В.Г. Царегородцев // Материалы X Всероссийского семинара "Нейроинформатика и ее приложения". – Красноярск, 2002. – С. 146 – 150.
8. Fisher R.A. The use of multiple measurements in taxonomic problems / R.A. Fisher // Annual Eugenics. – 1936. – Vol. 7. – Part II. – P. 179 – 188.
9. The plant recognition on remote sensing results by the feed-forward neural networks / V. Dubrovin, S. Subbotin, S. Morshchavka et al. // Smart Engineering System Design. – 2001. – N 3. – P. 251 – 256.
10. Кривенко В.И. Нейросетевое моделирование суммарного показателя качества жизни больных хроническим обструктивным бронхитом в ассоциации с клиническими особенностями течения заболевания / В.И. Кривенко, Л.Н. Евченко, С.А. Субботин // Вестник новых медицинских технологий. – 2001. – Т. VIII, № 4. – С. 7 – 10.

Стаття надійшла до редакції 08.08.2009