УДК 004.93

Субботин С. А.

Д-р техн. наук, профессор, заведующий кафедрой программных средств Запорожского национального технического университета, Запорожье, Украина

МЕТРИКИ КАЧЕСТВА ВЫБОРОК ДАННЫХ И МОДЕЛЕЙ ЗАВИСИМОСТЕЙ, ОСНОВАННЫЕ НА ФРАКТАЛЬНОЙ РАЗМЕРНОСТИ

Актуальность. Рассмотрена задача автоматизации формирования выборок из исходных выборок большого объема для построения моделей по прецедентам. Объектом исследования являлась модель качества выборки для построения моделей по прецедентам.

Цель работы – создание набора показателей для оценки качества выборок, имеющих единую природу, на основе принципов фрактального анализа.

Метод. Предложен комплекс показателей, позволяющих характеризовать качество подвыборок относительно исходной выборки с единых позиций на основе принципов фрактального анализа. Предложены методы определения фрактальной размерности выборки, оперирующие прямоугольными блоками одинакового размера, покрывая ними пространство признаков: не учитывающий характеристики синтезируемой модели, учитывающий ошибку (точность), синтезируемой модели, а также учитывающий точность и сложность синтезируемой модели. Наряду с фрактальной размерностью также предложен метод определения показателей качества выборки на основе принципа массовой размерности применительно к анализу данных. Предложенный метод разбивает пространство признаков на кластеры одинакового размера и формы. Варьируя размер кластера, метод позволяет получать различные уровни детализации выборки. Метод позволяет определить центр масс класса в выборке, среднее расстояние между экземплярами кластера, нормированное среднее отклонение расстояний между экземплярами от их среднего, массу и плотность экземпляров кластера, объем и площадь поверхности прямоугольного кластера, отношение объема к площади поверхности кластера, средневзвешенную равномерность расположения экземпляров в кластерах класса, массу и плотность экземпляров класса, средневзвешенную равномерность расположения экземпляров выборки.

Результаты. Разработанные показатели реализованы программно и исследованы при решении задачи классификации ирисов Фишера.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенного математического обеспечения и позволяют рекомендовать его для использования на практике при решении задач диагностирования и автоматической классификации по признакам. Перспективы дальнейших исследований могут заключаться в создании параллельных методов расчета комплекса предложенных показателей, оптимизации их программных реализаций, а также экспериментальном исследовании предложенных показателей на большем комплексе практических задач разной природы и размерности.

Ключевые слова: выборка, фрактальная размерность, метрика качества, кластер, формирование выборок.

НОМЕНКЛАТУРА

$\Gamma(x)$ – гамма-функция;

ε – граничное значение ошибки модели;

κ – определяемый пользователем параметр;

ξ^k – средневзвешенная равномерность расположения экземпляров в кластерах k-го класса;

ξ – средневзвешенная равномерность расположения экземпляров выборки;

ρ^k – плотность экземпляров k-го класса;

$\rho^{k,q}$ – плотность экземпляров q-го кластера k-го класса;

ρ_E – отношение, показывающее ожидание сложности отображения;

$\sigma^{k,q}$ – нормированное среднее отклонение расстояний между экземплярами q-го кластера k-го класса от их среднего;

$\upsilon_S^{k,q}$ – отношение объема шара, ограниченного гиперсферой, к площади поверхности гиперсферы q-го кластера k-го класса;

Ω_E – ожидание сложности нейронной сети прямого распространения;

Ω_i – сложность связей каждой субструктуры сети;

ω – множество настраиваемых параметров модели;

C^k – центр масс k-го класса в выборке;

C_j^k – j-я координата центра масс k-го класса в выборке;

D – фрактальная размерность выборки;

$D_{net(w)}$ – фрактальная размерность данных относительно точности (ошибки) E синтезированной модели net при текущем w;

D_{net} – фрактальная размерность данных относительно точности (ошибки) синтезированной модели;

D_{net*} – фрактальная размерность модели net**;**

D_E – фрактальная размерность ожидания сложности двухслойной сети прямого распространения;

$D^{(k)}$ – фрактальная размерность k-го класса;

$\langle Dc \rangle$ – корреляционная размерность;

D_C – фрактальная размерность сложности связей многослойной нейросети прямого распространения;

E – критерий качества обучения модели (функция ошибки);

e – число структурных элементов распознающей модели;

f – критерий качества модели;

$F()$ – структура модели;

h – число скрытых слоев сети;

H_i – i-й скрытый слой сети;

DOI 10.15588/1607-3274-2017-2-8

j – номер признака

K – число классов;

L – число интервалов, на которые разбиваются диапазоны значений признаков;

l – длина интервала признака или размер гиперкуба;

M – число выходов сети;

M^k – масса экземпляров k-го класса относительно центров масс его кластеров;

$M^{k,q}$ – масса экземпляров q-го кластера k-го класса;

N – число входных признаков или число входов сети;

$n_{(k)}$ – число гиперблоков со стороной размером l, покрывающих k-й класс выборки в пространстве N признаков;

$n(l)$ – число гиперблоков со стороной размером l, покрывающих выборку;

$net(w)$ распознающая модель;

$n_{i,q}$ – число экземпляров, попавших в прямоугольный гиперблок, образованный q-м интервалом i-го признака;

$n_{i,q,k}$ – число экземпляров k-го класса, попавших в прямоугольный гиперблок, образованный q-м интервалом i-го признака;

opt – условное обозначение оптимума;

$P_s^{k,q}$ – площадь поверхности гиперсферы, ограничивающей q-й кластер k-го класса;

q – номер кластера;

Q – число кластеров всех классов;

Q^k – число кластеров в k-м классе;

r – единичный радиус кластера;

R – множество расстояний, меньших r;

$|R|$ – мощность множества R;

$R^{k,q}(s,p)$ – расстояния между s-м и p-м экземплярами k-го класса в зоне q-го кластера, отстоящей от центра k-го класса не более чем на r;

$\overline{R}$ – среднее расстояние;

r – радиус отсечения (cut-off radius);

r_k – евклидово расстояние между парой точек;

S – число прецедентов;

S_E – размер входного и выходного слоев, определяющий структурную сложность модели;

S_M – размер выходного слоя нейронной сети прямого распространения;

S_N – размер входного слоя нейронной сети прямого распространения;

$S_{(H_i,H_{i+1})}$ – суммарное число нейронов в i-м и (i+1)-м слоях сети;

$V_s^{k,q}$ – объем шага, ограниченного гиперсферой q-го кластера k-го класса;

w – число параметров модели;

X – исходная выборка;

$<x^s, y^s>$ – s-й прецедент;

x_j – j-й признак;

x_j^s – значение j-го входного признака для s-го прецедента (экземпляра) выборки;

$x_{j,\max}^{k,q}$ – максимальное значение j-го признака для экземпляров принадлежащих к q-му кластеру k-го класса;

$x_j^{\max}, x_j^{\min}$ – соответственно, максимальное и минимальное значения признака x_j;

$x_{j,\min}^{k,q}$ – минимальное значение j-го признака для экземпляров принадлежащих к q-му кластеру k-го класса;

y^s – значение выходного признака для s-го прецедента (экземпляра) выборки.

ВВЕДЕНИЕ

При построении моделей принятия решений по прецедентам весьма важной задачей является формирование выборок данных, поскольку оно позволяет существенно ускорить процесс обучения модели путем выделения репрезентативной обучающей выборки малого объема.

Известные методы формирования выборок [1–5], как правило, представляют собой переборные стратегии и требуют задания в качестве целевой функции некоторого критерия, характеризующего качество формируемой подвыборки относительно исходной выборки большого объема.

Объектом исследования являлась модель качества выборки для построения моделей по прецедентам.

Ранее автором в [4, 6–9] был предложен комплекс показателей, образующий модель качества выборки. Данная модель позволяет характеризовать такие свойства выборки, как компактность, монотонность, нелинейность, отделимость классов, повторяемость, полнота, противоречивость, равномерность и неравномерность, размерность, разнообразие, репрезентативность, связанность переменных, сложность, эластичность и др.

Предметом исследования являлись показатели качества выборок.

Показатели качества выборок, входящие в модель [4, 6–9] имеют разную природу и не позволяют в едином ключе сравнивать качество выборок. С одной стороны, это влечет большие затраты времени и памяти на расчет всего комплекса показателей, а, с другой стороны, это вызывает сложности в объединении и совместном использовании комплекса показателей разной природы.

Поэтому актуальной является задача разработки показателей, позволяющих оценивать качество выборок с единой позиции.

Одним из перспективных направлений анализа данных является фрактальный анализ [10–15], ключевым понятием которого является фрактальная размерность – коэффициент, описывающий фрактальные структуры или множества на основе количественной оценки их сложности как коэффициент изменения в деталях с изменением масштаба.

Целью данной работы было создание набора показателей для оценки качества выборок, имеющих единую природу, на основе принципов фрактального анализа.

1 ПОСТАНОВКА ЗАДАЧИ

Пусть мы имеем исходную выборку $X = <x, y>$ – набор S прецедентов о зависимости $y(x)$, $x = \{x^s\}$, $y=\{y^s\}$, $s = 1, 2, ..., S$, характеризующихся набором N входных признаков $\{x_j\}$, $j = 1, 2, ..., N$, и выходным признаком y. Каждый s-й прецедент представим как $<x^s, y^s>$, $x^s=\{x_j^s\}$, $y^s \in \{1, 2, ..., K\}$, $K>1$.

Тогда задача синтеза модели зависимости $y(x)$ будет заключаться в определении таких структуры $F()$ и значений параметров w модели, при которых будет удовлетворен критерий качества модели $f(F(),\ \omega,\ <x,\ y>)\rightarrow opt$. Обычно критерий качества обучения нейросетей определяют как функцию ошибки модели:

$$E=\frac{1}{2}\sum_{s=1}^{S}(y^s-F(\omega,x^s))^2\rightarrow\min.$$

Для задач с дискретным выходом ошибку обученной модели можно характеризовать также формулой:

$$E=\frac{100\%}{S}\sum_{s=1}^{S}\left|y^s-F(\omega,x^s)\right|\rightarrow\min.$$

В случае, когда исходная выборка имеет большую размерность, перед построением нейромодели необходимо решить задачу выделения обучающей выборки меньшего объема: дано: $<x,\ y>$, надо: $<x',\ y'>,\ x'\in\{x^s\}$, $y'=\{y^s|x^s\in x'\}$, $S'=|y'|$, $S'<S, f(<x',\ y'>, <x,\ y>)\rightarrow opt$.

Приведенная задача требует определения критерия f, позволяющего отображать соответствие свойств формируемой подвыборки свойствам исходной выборки.

2 ЛИТЕРАТУРНЫЙ ОБЗОР

Согласно [16] размерность Хаусдорфа-Бесиковича (Hausdorff – Besicovich dimension) определяется как

$$D\approx\frac{\log(n(l))}{\log\left(\frac{1}{l}\right)},$$

где $n(l)$ – минимальное число гиперкубов размером l, необходимых для покрытия образа.

Одним из наиболее доступным способов определения размерности Хаусдорфа-Бесиковича является метод подсчета (box-counting method) [14, 15], заключающийся в повторяющемся покрытии фрактального объекта гиперкубами равного размера и подсчетом каждый раз минимального числа гиперкубов, которые содержат точки образа.

Последовательно уменьшая размер гиперкубов l, получают набор точек с координатами $(\log(n(l)),\ \log\left(\frac{1}{l}\right))$, задающий кривую, наклон которой, определяемый с помощью линейной регрессии, является фрактальной размерностью:

$$D=\lim_{l\rightarrow 0}\frac{\log(n(l))}{\log\left(\frac{1}{l}\right)}.$$

Метод Такенса (Takens' method) [17, 18] используется для определения корреляционной размерности

$$\left\langle Dc\right\rangle=-\left\{\frac{1}{|R|}\sum_{k=1}^{|R|}r_k\right\}^{-1},$$

где $R=\{r_k|\ r_k<r\}$, $r>0$, r задается эвристически.

Общим недостатком рассмотренных методов определения фрактальной размерности [14–18] является то, что мощность множества должна удовлетворять неравенству $N<2\log_{10}S$, показывающему, что число точек данных S, необходимых для точного оценивания размерности N-мерного множества должно быть как минимум $10^{\frac{N}{2}}$. Даже для множеств малого объема это приводит к большим значениям N.

Также фрактальную размерность в [19] предлагается характеризовать соотношениями «масса-радиус», «периметр-площадь», «площадь-объем». Однако предложенные показатели [19] определены для двумерных графических изображений и неприменимы для многомерных выборок данных.

Для нейросетевых моделей, обученных по прецедентам, в [20] предложен ряд способов определения фрактальной размерности.

Фрактальная размерность ожидания сложности двухслойной сети прямого распространения определяется как

$$D_E=\frac{\log(\Omega_E)}{\log\left(\frac{1}{\gamma_E}\right)},$$

где $\Omega_E=(1+\rho_E{}^{\kappa})S_E$, $\rho_E=\frac{\max(S_N,S_M)}{\min(S_N,S_M)}$, $S_E=S_N+S_M$, $\kappa>0$, $\frac{1}{\gamma_E}=1+\frac{1}{\rho_E}$ – соотношение, используемое для задания собственной шкалы измерений, принимая во внимание возможность $\rho_E=1$ такую, что $\log(\gamma_E)=0$, если $\gamma_E=\rho_E$.

Фрактальная размерность сложности связей многослойной нейросети прямого распространения D_C определяется как сумма фрактальных размерностей всех субструктур сети:

$$D_C=\sum_{i=1}^{h+1}\frac{\log\Omega_i}{\log\left(\frac{1}{\gamma_i}\right)},$$

где $1\le i\le h+1$: $\Omega_i=(1+\mathrm{P}_i^{\kappa})S_{(H_i,H_{i+1})}$,

$\mathrm{P}_i=\frac{\sum\left\{S_{(H_j)}\,|\,H_j\in\Omega_i\right\}}{N+M}$, γ_i изменяется на $\frac{1}{\gamma_i}=1+\frac{1}{\mathrm{P}_i}$ с учетом возможности $\mathrm{P}_i=1$, так, что $\log(\gamma_i)=0$, если $\gamma_i=\mathrm{P}_i$.

Для создания структуры нейросети фрактальная размерность сложности связей D_c не может быть меньше ее фрактальной размерности ожидания сложности D_E, т.е. $D_c\ge D_E$. Когда $D_c\approx D_E$ установленная структура сети может считаться оптимальной.

Особенностью рассмотренных известных методов определения фрактальной размерности является то, что

размерность выборки и размерность модели, построенной на ее основе, определяются без связи друг с другом. Это ограничивает их практическое применение. Поэтому необходимо разработать методы оценивания фрактальной размерности, позволяющие характеризовать свойства выборки и свойства обученной модели по выборке.

3 МАТЕРИАЛЫ И МЕТОДЫ

Экземпляры выборки можно представить как точки в пространстве признаков. Тогда кластеры будут соответствовать компактным областям в пространстве признаков, которые будут объединяться в классы. Кластеры можно описывать различными геометрическими фигурами. Фрактальный анализ выборки в пространстве признаков можно осуществить, задав элементарную форму для выделения кластеров и варьируя размер кластера для разбиения выборки на фрагменты.

Для анализа фрактальной размерности выборки предлагается использовать следующий метод.

Этап инициализации. Задать обучающую выборку $<x,y>$ и L.

Этап нормирования выборки. Если значения признаков ненормированы, то их следует пронормировать, отобразив на интервал [0, 1]:

$$x_j^s = \frac{x_j^s - x_j^{\min}}{x_j^{\max} - x_j^{\min}}.$$

Этап кластеризации. Разбить диапазон значений каждого признака на L интервалов длиной l:

$$l = \frac{1}{L}.$$

Сформировать кластеры как прямоугольные блоки на пересечении интервалов разных признаков.

Этап анализа данных. Определить число экземпляров, попавших в каждый прямоугольный гиперблок, образованный интервалами признаков $n_{i,q}$

Определить число экземпляров k-го класса, попавших в каждый прямоугольный гиперблок, образованный интервалами признаков $n_{i,q,k}$

Определить число гиперблоков со стороной размером l, покрывающих k-й класс выборки в пространстве N признаков

$$n_{(k)} = \sum_{i=1}^{N}\sum_{q=1}^{L}\left\{1 \mid n_{i,q,k} > 0\right\}.$$

Определить число гиперблоков со стороной размером l, покрывающих выборку в пространстве N признаков

$$n(l) = \sum_{i=1}^{N}\sum_{q=1}^{L}\left\{1 \mid n_{i,q} > 0\right\} = \sum_{i=1}^{N}\sum_{q=1}^{L}\left\{1 \left| \sum_{k=1}^{K} n_{i,q,k} > 0 \right.\right\}.$$

Этап определения фрактальной размерности. Определить при заданном l фрактальную размерность k-го класса, k= 1, 2, ..., K:

$$D^{(k)} = \frac{\log(n_{(k)})}{\log\left(\frac{1}{l}\right)}.$$

Определить фрактальную размерность выборки при заданном l:

$$D = \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)}.$$

Данный метод оперирует прямоугольными блоками одинакового размера, покрывая ними пространство признаков. Единственным управляемым параметром метода является задаваемое число интервалов L, на которые разбиваются диапазоны значений признаков. Очевидно, что число кластеров $Q \geq K$, а $Q=L^N$, причем для каждого признака $L \geq 2$. Для обеспечения обобщающих свойств кластеров введем ограничение $Q \leq NS$. Таким образом, получим: $K \leq L^N \leq NS$, $L \geq 2$. Прологарифмируем: $\log(K) \leq N\log(L) \leq \log(NS)$ и после преобразований получим: $2L$. Заметим, что минимальным шагом для варьирования значений L является 1. Если значение верхнего предела оказывается меньше 2, то его можно заменить на S. Это связано с тем, что на оси каждого признака будет не более чем S точек и разбиение их на более чем S интервалов, очевидно, будет приводить к возникновению пустых интервалов. При больших значениях N задание числа разбиений порядка S по каждому признаку приведет к выделению огромного числа блоков порядка S^N, что сделает вычисления крайне трудоемкими, а в ряде случаев и практически не реализуемыми. Поэтому рационально в этом случае ограничить значение верхнего предела значений L величиной порядка round(lg(S)), где roung – функция округления к ближайшему целому числу.

Определение показателя D при малых значениях L потребует больших затрат вычислительных ресурсов и ресурсов памяти ЭВМ, чем при больших значениях L. Однако точность анализа для малых значений L будет ниже, в то время как уровень обобщения будет выше, чем для больших значений L.

Достоинством предложенного метода и определяемого на его основе показателя качества выборки является то, что они не зависят от метода синтеза модели и результатов его работы и позволяют оценивать свойства отдельно взятой выборки.

Недостатками предложенного метода является неопределенность в выборе значения параметра L, а также отсутствие связи метода с качеством синтезируемой модели.

Для устранения отмеченных недостатков предлагается использовать метод определения фрактальной размерности входного множества для обучения распознающей модели.

Этап инициализации. Задать обучающую выборку $<x,y>$, метод синтеза модели, критерий качества обучения модели как функцию ошибки E, а также максимально приемлемое значение ошибки ε.

Этап нормирования выборки. Если значения признаков ненормированы, то их следует пронормировать, отобразив на интервал [0, 1].

Этап формирования и анализа разбиения данных. Последовательно изменяя значение L = 2, ..., S:

– определить длину интервала l;

– квантовать признаки выборки, разбив диапазоны их значений на L интервалов;

– определить число гиперблоков со стороной размером l, покрывающих выборку в пространстве N признаков, $n(l)$;

построить распознающую модель *net* с помощью заданного метода, минимизируя функцию ошибки E до достижения приемлемого уровня ε;

оценить ошибку E построенной распознающей модели *net*.

Этап определения фрактальной размерности. Для всех l, для которых ошибка модели является приемлемой, определить фрактальную размерность данных относительно точности (ошибки) синтезированной модели:

$$D_{net} = \left\{ \left| \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)} \right| E(net) \le \varepsilon \right\}.$$

Данный метод оперирует прямоугольными блоками одинакового размера, покрывая ними пространство признаков. Единственным управляемым параметром метода является задаваемое граничное значение ошибки ε модели. Очевидно, что чем меньше будет заданное ε, тем более детальной должна быть модель, то есть потребуется выделить большее число кластеров Q, а, следовательно, и большим будет число L. Соответственно, с уменьшением заданного ε будут возрастать затраты вычислительных ресурсов и ресурсов памяти ЭВМ на анализ выборки.

Достоинством предложенного метода и определяемого на его основе показателя качества выборки является то, что они связаны с показателем качества синтезируемой модели, а также в автоматическом режиме устанавливает оптимальное значение параметра L.

Недостатками предложенного метода является неопределенность в выборе значения параметра ε, а также его зависимость от качества обучения и принципов функционирования модели, по которой он определяется. Также следует отметить, что функция ошибки, используемая в методе, является лишь одной из характеристик синтезируемой модели, но не учитывает ее размерность и обобщающие свойства.

Поэтому фрактальную размерность обученной модели предлагается определять на основе изложенного ниже метода, учитывающего размерность модели.

Размерность модели – число величин, представляющих в модели параметры и характеристики. Для распознающей модели оценим размерность числом ее параметров w и числом структурных элементов e. Для разных типов моделей эти величины могут существенно отличаться. Однако, как правило, w зависит от e, причем $w>>e$. Поэтому предлагается для оценки размерности распознающей модели ограничимся числом ее параметров w.

Определим максимальное допустимое с точки зрения обобщения значение w как размерность входа обучающего множества $w_{max} = NS$, а минимально допустимое значение $w_{min} > 0$.

Этап инициализации. Задать обучающую выборку $<x,y>$, метод синтеза модели, критерий качества обучения модели как функцию ошибки E, а также максимально приемлемое значение ошибки ε. Задать $w = w_{max.}$

Этап нормирования выборки. Если значения признаков ненормированы, то их следует пронормировать, отобразив на интервал [0, 1].

Этап формирования и анализа разбиения данных. Последовательно изменяя значение $L = 2, ..., S$:

– определить длину интервала l;

– квантовать признаки выборки, разбив диапазоны их значений на L интервалов;

– определить число гиперблоков со стороной размером l, покрывающих выборку в пространстве N признаков, $n(l)$;

построить распознающую модель $net(w)$ с помощью заданного метода, минимизируя функцию ошибки E до достижения приемлемого уровня ε;

оценить ошибку E построенной распознающей модели *net*.

Этап минимизации сложности модели. Найти минимальное l при котором точность построенной модели $net(w)$ является приемлемой. Последовательно изменяя значения $w = w_{min}, ..., w_{max}$ построить модель $net(w)$ с помощью заданного метода, минимизируя функцию ошибки E до достижения приемлемого уровня ε.

Для всех w, для которых ошибка модели является приемлемой, определить фрактальную размерность данных относительно точности (ошибки) синтезированной модели при текущем w:

$$D_{net(w)} = \left\{ \left| \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)} \right| E(net(w)) \le \varepsilon \right\}.$$

Для $D_{net(w)} > 0$ определить такое минимальное w, при котором при минимальном L ошибка модели E будет приемлемой:

$$D_{net*} = \left\{ \left| \frac{D_{net(w)}}{\log\left(\frac{1}{w}\right)} \right| E(net(w)) \le \varepsilon \right\} =$$

$$= \left\{ \left| \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)\log\left(\frac{1}{w}\right)} \right| E(net(w)) \le \varepsilon \right\}.$$

Данный метод оперирует прямоугольными блоками одинакового размера, покрывая ними пространство признаков. Управляемыми параметрами метода является задаваемое граничное значение ошибки ε модели, а также число параметров модели w. Очевидно, что чем меньше будет значение w, тем выше может быть уровень обобщения модели, однако и тем будет сложнее построить модель с требуемой точностью ε.

Наряду с фрактальной размерностью также возможно рассмотреть принцип массовой размерности применительно к анализу данных.

Рассмотрим разбиение признакового пространства на компактные области – кластеры одинакового размера и формы. Каждый кластер будет содержать близко расположенные экземпляры, обладающие подобными характеристиками (значениями описательных признаков). Очевидно, что варьируя размер кластера, мы получим различные уровни детализации выборки. Определим допустимый диапазон варьирования числа кластеров. Поскольку каждый класс может быть представлен в пространстве признаков не менее, чем одним кластером, то нижняя граница числа кластеров не может быть меньше числа классов. С другой стороны, при числе кластеров большем, чем число экземпляров в выборке, получим большое число пустых кластеров. Таким образом, $K \le Q \le S$.

Также, очевидно, что размерность описания кластеров не должна превышать размерность исходной выборки. Размерность исходной выборки оценим как NS, а каждый кластер будет описываться центром в N-мерном пространстве признаков и радиусом по оси каждого признака. Если считать все признаки предварительно пронормированными, а кластеры – одного размера и формы, то описание кластера будет иметь размерность N+1. Следовательно, $(N+1)Q \le NS \Rightarrow Q \le NS/(N+1)$. Поскольку N>1, то $K \le Q \le S \le NS/(N+1)$.

Очевидно, что при большом числе экземпляров в исходной выборке придется рассматривать большое число кластеров, что приведет к существенным затратам времени. Поэтому предлагается метод, позволяющий выделять минимизировать число анализируемых кластеров, последовательно наращивая их по необходимости.

Этап инициализации. Задать выборку <x, y>. Задать единичный радиус r.

Этап анализа классов. Для каждого класса k=1,2,...,K:

– найти центр масс k-го класса $C^k=\{C^k_j\}$ среди всех имеющихся в выборке экземпляров данного k-го класса:

$$C^k_j = \frac{1}{S^k}\sum_{s=1}^{S}\left\{x^s_j \middle| y^s = k\right\},\ j = 1,2,...,N;$$

– установить номер текущего кластера k-го класса q=1.

– выполнить этап анализа q-го кластера.

Этап анализа q-го кластера:

– если текущий кластер оказался пустым (не содержащим экземпляров k-го класса), то принять в качестве центра текущего кластера самый удаленный от текущего центра экземпляр того же класса;

– в зоне, отстоящей от центра k-го класса не более чем на r, найти расстояния между экземплярами соответствующего класса:

$$R^{k,q}(s,p) = \sqrt{\sum_{j=1}^{N}\left\{\left(x^s_j - x^p_j\right)^2 \middle| R(x^s, C^{k,q}) \le r, R(x^p, C^{k,q}) \le r, y^s = y^p = k\right\}},$$

$$R^{k,q}(p,s) = R^{k,q}(s,p),$$

$$s = 1, 2, ..., S, p = s+1, s+2, ..., S,$$

где

$$R(x^s, C^{k,q}) = \sqrt{\sum_{j=1}^{N}\left(x^s_j - C^{k,q}_j\right)^2},$$

$$C^{k,q}_j = C^{k,q}_j;$$

– определить среднее расстояние:

$$\overline{R} = \frac{1}{S(S-1)}\sum_{s=1}^{S}\left\{\sum_{p=s+1}^{S}\left\{R^{k,q}(s,p) \middle| R(x^p, C^{k,q}) \le \right.\right.$$

$$\left.\left. \le r, y^p = k\right\} \middle| R(x^s, C^{k,q}) \le r, y^s = k\right\};$$

– определить нормированное среднее отклонение расстояний между экземплярами от их среднего:

$$\sigma^{k,q} = \frac{1}{r}\sqrt{\sum_{s=1}^{S}\left\{\sum_{p=1}^{S}\left\{\left(R^{k,q}(s,p) - \overline{R}\right)^2 \middle| R(x^p, C^{k,q}) \le r, y^p = k\right\} \middle| R(x^s, C^{k,q}) \le r, y^s = k\right\}}.$$

Значение показателя будет находиться в диапазоне [0, 1]. Чем меньше будет значение показателя, тем равномернее расположены экземпляры класса в кластере;

– определить массу экземпляров кластера:

$$M^{k,q} = \sum_{s=1}^{S}\left\{\frac{1}{1 + R(x^s, C^{k,q})}\right\};$$

– определить плотность экземпляров кластера:

$$\rho^{k,q} = \frac{M^{k,q}}{S^{k,q}};$$

– определить площадь поверхности гиперсферы размерности N для q-го кластера k-го класса:

$$P^{k,q}_s = NC_N\left(\frac{1}{2}\max_{j=1,2,...,N}\{x^{k,q}_{j,\max} - x^{k,q}_{j,\min}\}\right)^{N-1},$$

где

$$C_N = \frac{\pi^{\frac{N}{2}}}{\Gamma\left(\frac{N}{2}+1\right)} = \begin{cases} \dfrac{\pi^{\frac{N}{2}}}{\left(\frac{N}{2}\right)!}, \text{если } N - \text{четное}; \\ \dfrac{2^{\frac{N-1}{2}+1}\,\pi^{\frac{N-1}{2}}}{N!!}, \text{если } N - \text{нечетное}, \end{cases}$$

– определить объем N-мерного шара, ограниченного гиперсферой размерности N для q-го кластера k-го класса:

$$V_s^{k,q} = C_N \left(\frac{1}{2} \max_{j=1,2,...,N} \{ x_{j,\max}^{k,q} - x_{j,\min}^{k,q} \} \right)^N ;$$

– определить отношение объема к площади поверхности кластера:

$$\upsilon_s^{k,q} = \frac{V_s^{k,q}}{P_s^{k,q}} ;$$

– удалить из рассмотрения экземпляры соответствующего q-го кластера k-го класса. Если среди оставшихся экземпляров в выборке все еще имеются экземпляры k-го класса, то положить $q=q+1$, скорректировать значение S^k, принять:

$$C_j^k = \frac{1}{S^k} \sum_{s=1}^{S} \left\{ x_j^s \middle| y^s = k \right\}, j=1, 2, ..., N,$$

перейти к этапу анализа q-го кластера; в противном случае – вернуть исходное значение S^k и перейти к этапу анализа выборки.

Этап анализа выборки. Для k=1, 2, ..., K определить:

– средневзвешенную равномерность расположения экземпляров в кластерах класса

$$\xi^k = \frac{1}{S^k} \sum_{q=1}^{Q^k} S^{k,q} \sigma^{k,q} .$$

Значение показателя будет находиться в диапазоне [0, 1]. Чем меньше будет значение показателя, тем равномернее расположены экземпляры класса в его кластерах;

– массу экземпляров класса относительно центров масс его кластеров:

$$M^k = \sum_q \left\{ \frac{M^{k,q}}{1 + R(C^{k,q}, C^k)} \right\},$$

где $R(C^{k,q}, C^k) = \sqrt{\sum_{j=1}^{N} \left(C_j^{k,q} - C_j^k \right)^2}$;

– плотность экземпляров класса:

$$\rho^k = \frac{M^k}{S^k} .$$

После чего определить средневзвешенную равномерность расположения экземпляров выборки:

$$\xi = \frac{1}{S} \sum_{k=1}^{K} S^k \rho^k = \frac{1}{S} \sum_{k=1}^{K} M^k .$$

Значение показателя будет находиться в диапазоне [0, 1]. Чем меньше будет значение показателя, тем равномернее расположены экземпляры класса в его кластерах.

Предложенный метод позволяет определить комплекс показателей, характеризующие свойства кластеров, классов и выборки в целом.

4 ЭКСПЕРИМЕНТЫ

Для исследования комплекса предложенных показателей выборок и моделей, основанных на фрактальной размерности, они были программно реализованы. Разработанное программное обеспечение использовалось для проведения вычислительных экспериментов по исследованию применимости предложенных показателей на примере решения задачи автоматической классификации ирисов Фишера по признакам [21], выборка данных для которой визуализирована на рис. 1.

5 РЕЗУЛЬТАТЫ

В результате проведенных экспериментов для задачи классификации ирисов Фишера оценка фрактальной размерности выборки составила D = 0,59034, а оценки фрактальных размерностей классов: $D^{(1)}$ = 0,68223, $D^{(2)}$ = 0,6212, $D^{(3)}$ = 0,53407. Графики зависимостей от l^{-1} в логарифмической системе координат для всей выборки и классов приведены на рис. 2а и 2б, соответственно. На рис. 2б разные классы кодируются маркерами разных размеров.

На рис. 3 приведены схематические графики зависимостей $n(l)$ выборки от l^{-1} в логарифмической системе координат для разных заданных значений ε и получаемых значений E (рис. 3а), а также размерностей формируемой выборки S' и N' (рис. 3б).

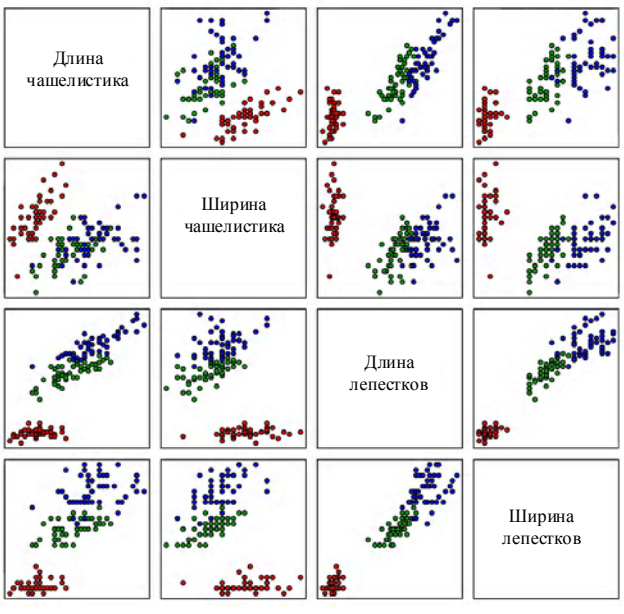
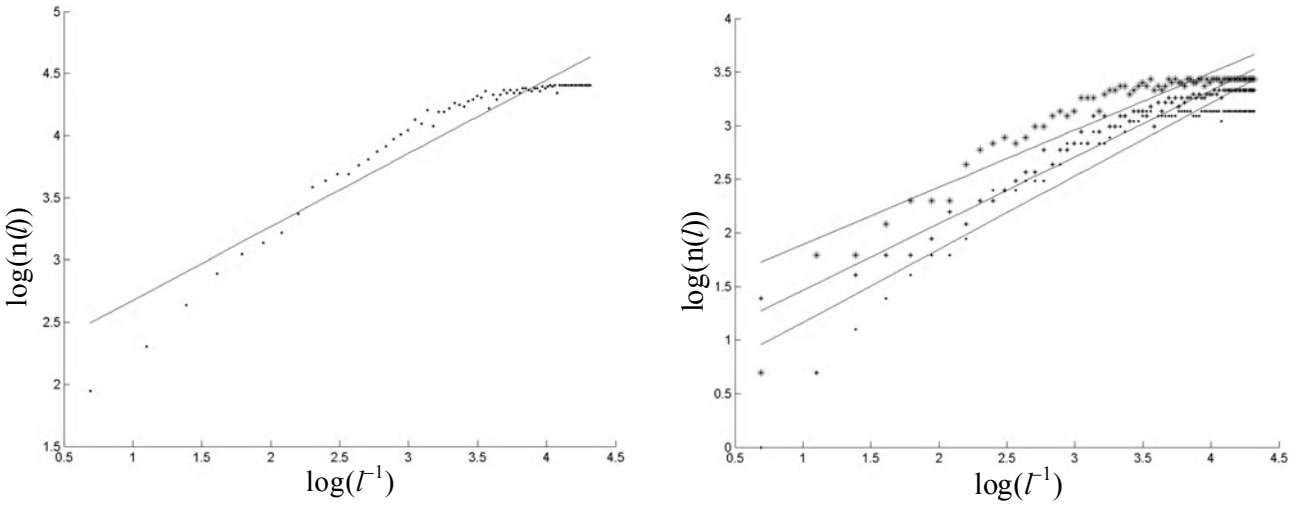
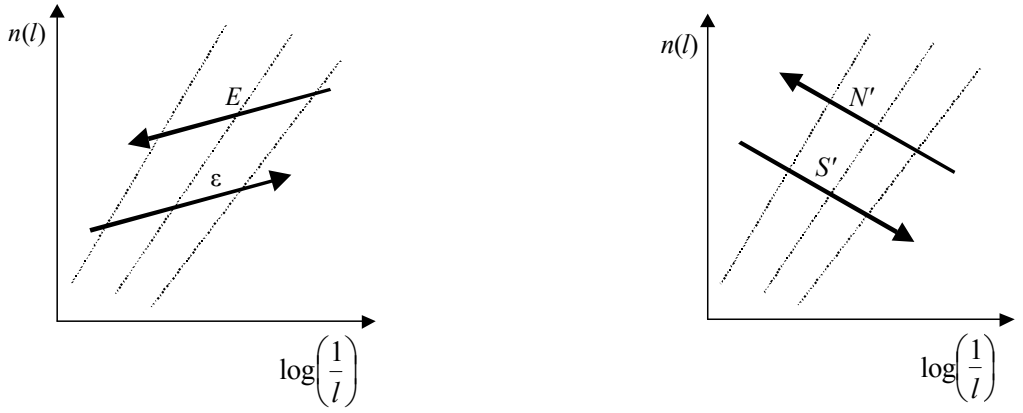


Рисунок 1 – Визуализация выборки данных ирисов Фишера для разных классов: красный – Ирис щети́нистый (Iris setosa), зеленый – Ирис разноцветный (Iris versicolor), синий – Ирис виргинский (Iris virginica) (Источник – https://upload.wikimedia.org/wikipedia/commons/thumb/5/56/Iris_dataset_scatterplot.svg/1024px-Iris_dataset_scatterplot.svg.png)



а б

Рисунок 2 – Графики зависимостей от l^{-1} в логарифмической системе координат:
а – $n(l)$ выборки, б – $n(l)$ классов



а б

Рисунок 3 – Схематические графики зависимостей от l^{-1} в логарифмической системе координат:
а – $n(l)$ выборки, б – $n(l)$ классов

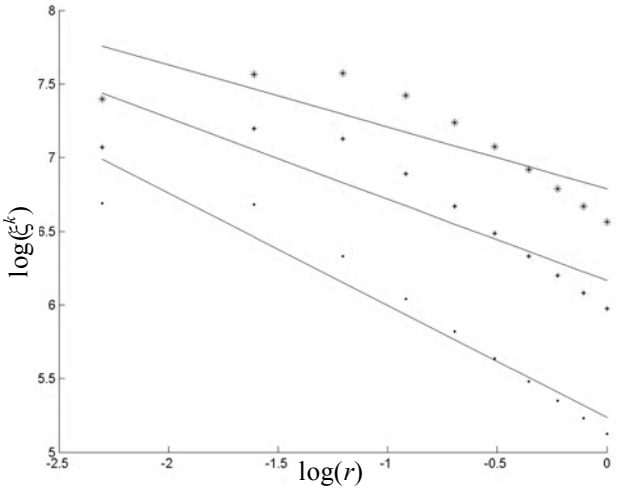
На рис. 4 изображены графики зависимостей: ξ^k от r (рис. 4а), ρ^k от r (рис. 4б), M^k от r (рис. 4в), υ^k от r (рис. 4г), ξ от r (рис. 4д) в логарифмической системе координат. На рис. 4а–рис. 4г разные классы кодируются маркерами разных размеров. В табл. 1 приведены рассчитанные фрактальные размерности классов и выборки, зависящие от r.

6 ОБСУЖДЕНИЕ

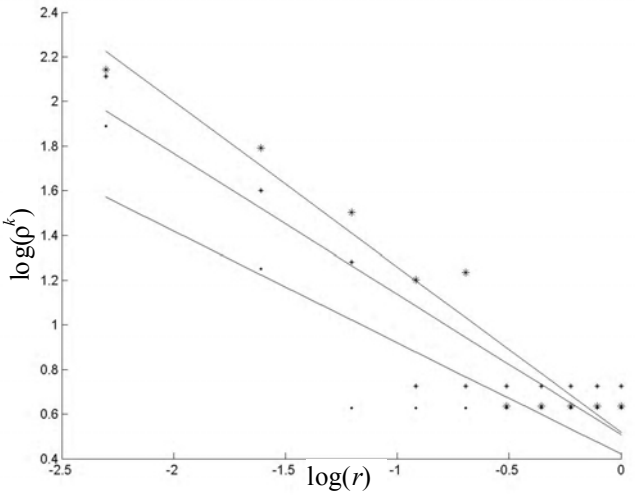
Проведенные эксперименты подтвердили работоспособность предложенных методов и реализующих их программных средств.

Как видно из рис. 2 и рис. 4, а также табл. 1, предложенные показатели фрактальной размерности на основе подсчета попаданий в блоки и соотношений массарадиус позволяют хорошо показать различия между классами. Эти показатели можно использовать в методах формирования выборок, определяя критерии качества формируемых подвыборок на основе предложенных показателей фрактальной размерности. Если формируемая подвыборка или ее классы по показателям фрактальной размерности существенно отличаются от аналогичных показателей исходной выборки, то возможно, что полученная выборка не обладает репрезентативностью относительно исходной выборки. Также при сравнении нескольких подвыборок-кандидатов предложенные показатели могут использоваться как меры их качества: из подвыборок-кандидатов следует отдавать предпочтение той, которая будет иметь показатели фрактальной размерности, наиболее близкие по значениям к показателям исходной выборки.

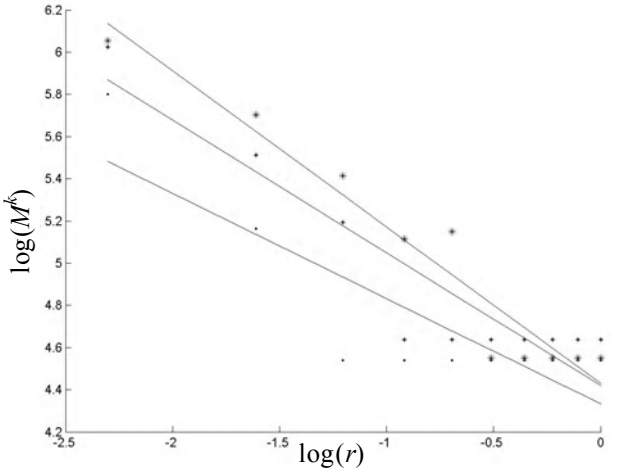
Как видно из рис. 3, изменение заданных составляющих размерности формируемой подвыборки (числа признаков N' и числа экземпляров S'), а также E и ε влияют на положение прямой, соединяющей точки зависимости $n(l)$.



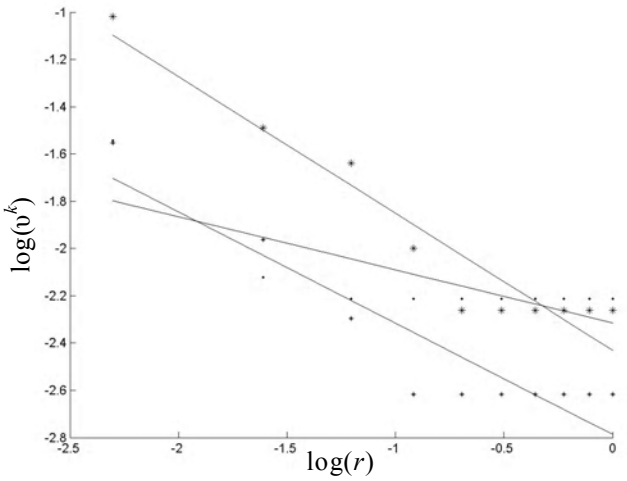
а



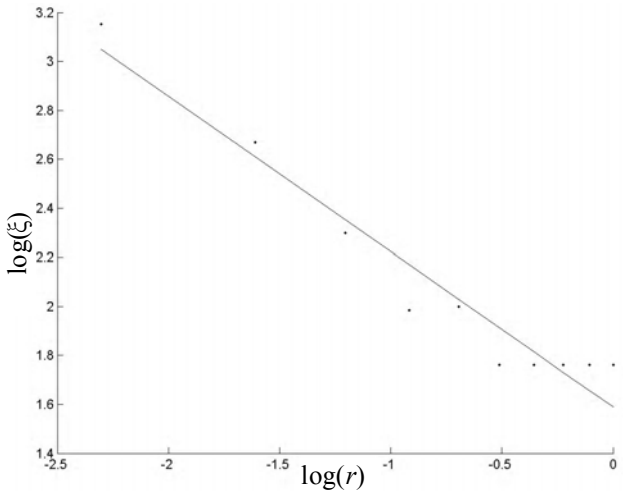
б



в



г



д

Рисунок 4 – Графики зависимостей от r в логарифмической системе координат:

а – ξ^k ; б – ρ^k ; в – M^k; г – υ^k ; д – ξ

Таблица 1 – Фрактальные размерности классов и выборки

Показатель фрактальной размерности	Метка класса			Выборка
	1	2	3	
ξ (ξ^k)	–0,7606	–0,5530	–0,4203	–0,6338
ρ^k	–0,4990	–0,6290	–0,7407	–
M^k	–0,4990	–0,6290	–0,7407	–
υ^k	–0,2243	–0,4715	–0,5789	–

ЗАКЛЮЧЕНИЕ

В работе рассмотрена актуальная задача автоматизации формирования выборок из исходных выборок большого объема для построения моделей по прецедентам.

Научная новизна полученных результатов заключается в том, что впервые предложен комплекс показателей, позволяющих характеризовать качество подвыборок относительно исходной выборки с единых позиций на основе принципов фрактального анализа.

Предложены методы определения фрактальной размерности выборки, оперирующие прямоугольными блоками одинакового размера, покрывая ними пространство признаков: не учитывающий характеристики синтезируемой модели, учитывающий ошибку (точность), синтезируемой модели, а также учитывающий точность и сложность синтезируемой модели.

Наряду с фрактальной размерностью также предложен метод определения показателей качества выборки на основе принципа массовой размерности применительно к анализу данных. Предложенный метод разбивает пространство признаков на кластеры одинакового размера и формы. Варьируя размер кластера, метод позволяет получать различные уровни детализации выборки. Метод позволяет определить центр масс класса в выборке, среднее расстояние между экземплярами кластера, нормированное среднее отклонение расстояний между экземплярами от их среднего, массу и плотность экземпляров кластера, объем и площадь поверхности прямоугольного кластера, отношение объема к площади поверхности кластера, средневзвешенную равномерность расположения экземпляров в кластерах класса, массу и плотность экземпляров класса, средневзвешенную равномерность расположения экземпляров выборки.

Практическая ценность полученных результатов состоит в том, что разработанные показатели реализованы программно и исследованы при решении задачи классификации ирисов Фишера.

Проведенные эксперименты подтвердили работоспособность предложенного математического обеспечения и позволяют рекомендовать его для использования на практике при решении задач построения моделей по прецедентам.

Перспективы дальнейших исследований могут заключаться в создании параллельных методов расчета комплекса предложенных показателей, оптимизации их программных реализаций, а также экспериментальном исследовании предложенных показателей на большем комплексе практических задач разной природы и размерности.

БЛАГОДАРНОСТИ

Работа выполнена в рамках госбюджетной научно-исследовательской темы «Методы и средства вычислительного интеллекта и параллельного компьютинга для обработки больших объемов данных в системах диагностирования» (номер гос. регистрации 0116U007419) кафедры программных средств Запорожского национального технического университета при частичной поддержке международного образовательного проета «Центры передового опыта для молодых ученых» (Ref. No. 544137-TEMPUS-1-2013-1-SK-TEMPUS-JPHES) программы «Темпус» Европейского Союза.

СПИСОК ЛИТЕРАТУРЫ

1. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken : John Wiley & Sons, 2008. – 339 p.
2. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York : Chapman & Hall, 2005. – 416 p. DOI: 10.1201/9781420028638
3. Encyclopedia of survey research methods / ed. P. J. Lavrakas. – Thousand Oaks: Sage Publications, 2008. – Vol. 1–2. – 968 p. DOI: 10.4135/9781412963947.n159
4. Субботин С. А. Формирование выборок и анализ качества моделей на основе нейронных и нейро-нечетких сетей в задачах диагностики и распознавания образов : монография / С. А. Субботин. – Saarbrьcken : LAP Lambert academic publishing, 2012. – 232 с. – (ISBN 978-3-8473-4471-1).
5. Кокрен У. Методы выборочного исследования / У. Кокрен ; пер. с англ. И. М. Сонина ; под ред. А. Г. Волкова, Н. К. Дружинина. – М. : Статистика, 1976. – 440 с.
6. Subbotin S. A. The training set quality measures for neural network learning / S. A. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2010. – Vol. 19, № 2. – P. 126–139. DOI: 10.3103/s1060992x10020037
7. Субботин С. А. Комплекс характеристик и критериев сравнения обучающих выборок для решения задач диагностики и распознавания образов / С. А. Субботин // Математичні машини і системи. – 2010. – № 1. – С. 25–39.
8. Субботин С. А. Критерии индивидуальной информативности и методы отбора экземпляров для построения диагностических и распознающих моделей / С. А. Субботин // Біоніка інтелекту. – 2010. – № 1. – С. 38–42.
9. Субботин С. А. Методы формирования выборок для построения диагностических моделей по прецедентам / С. А Субботин // Вісник Національного технічного університету «Харківський політехнічний інститут» : зб. наук. праць. – Харків : НТУ «ХПІ», 2011. – № 17. – С. 149–156.
10. Roberts A. Unbiased estimation of multi-fractal dimensions of finite data sets / A. Roberts, A. Cronin // Physica A: Statistical Mechanics and its Applications. – 1996. – Vol. 233, № 3–4. – P. 867-878. DOI:10.1016/s0378-4371(96)00165-3

11. Evaluating the fractal dimension of profiles / [B. Dubuc, J. Quiniou, C. Roques-Carmes, C. Tricot, S. Zucker] // Physical Review. – 1989. – T. 39, № 3. – P. 1500–1512. DOI:10.1103/PhysRevA.39.1500
12. Cheng Q. Multifractal Modeling and Lacunarity Analysis / Q. Cheng // Mathematical Geology. – 1997. – Vol. 29, № 7. – P. 919–932. DOI:10.1023/A:1022355723781
13. Eftekhari A. Fractal Dimension of Electrochemical Reactions / A. Eftekhari // Journal of the Electrochemical Society. – 2004. – Vol. 151, № 9. – P. E291–E296. DOI:10.1149/1.1773583.
14. Signal attenuation and box-counting fractal analysis of optical coherence tomography images of arterial tissue / [D. P. Popescu, C. Flueraru, Y. Mao at al] // Biomedical Optics Express. – 2010. – Vol. 1, № 1. – P. 268–277. DOI:10.1364/boe.1.000268
15. Li J. An improved box-counting method for image fractal dimension estimation / J. Li , Q. Du, C. Sun // Pattern Recognition. – 2009. – Vol. 42, № 11. – P. 2460–2469. DOI:10.1016/j.patcog.2009.03.001.
16. Crişan D. A. Fractal dimension spectrum as an indicator for training neural networks / D. A. Crişan, R. Dobrescu // Universitatea Politehnica Bucuresti Sci. Bull. Series C. – 2007. – Vol. 69, № 1. – P. 23–32.
17. Camastra F. Data Dimensionality Estimation Methods: A survey / F. Camastra // Pattern Recognition. – 2003. – Vol. 36, Issue 12. – P. 2945–2954. DOI: 10.1016/S0031-3203(03)00176-6
18. Takens F. On the numerical determination of the dimension of an attractor / F. Takens // Dynamical Systems and Bifurcations : Workshop, Groningen, 16–20 April 1984 : proceedings / [eds.: Braaksma B., Broer H. W., Takens F.]. – Berlin : Springer, 1985. – P. 99–106. – (Lecture Notes in Mathematics , Vol. 1125). DOI: 10.1007/bfb0075637
19. Чумак О. В. Энтропии и фракталы в анализе данных / О. В. Чумак. – М.-Ижевск : НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2011. – 164 с.
20. Zong-Chang Y. Establishing structure for artificial neural networks based-on fractal / Y. Zong-Chang // Journal of Theoretical and Applied Information Technology. – 2013. – Vol. 49, № 1. – P. 342–347.
21. Fisher Iris dataset [Electronic resource]. – Access mode: https://archive.ics.uci.edu/ml/datasets/Iris

Статья поступила в редакцию 25.01.2017.
После доработки 06.02.2017.

Субботін С. О.
Д-р техн. наук, професор, завідувач кафедри програмних засобів Запорізького національного технічного університету, Запоріжжя, Україна

МЕТРИКИ ЯКОСТІ ВИБІРОК ДАНИХ І МОДЕЛЕЙ ЗАЛЕЖНОСТЕЙ, ЗАСНОВАНІ НА ФРАКТАЛЬНІЙ РОЗМІРНОСТІ

Актуальність. Розглянуто задачу автоматизації формування вибірок з вихідних вибірок великого обсягу для побудови моделей за прецедентами. Об'єктом дослідження є модель якості вибірки для побудови моделей за прецедентами.

Мета роботи – створення набору показників для оцінки якості вибірок, що мають єдину природу, на основі принципів фрактального аналізу.

Метод. Запропоновано комплекс показників, що дозволяють характеризувати якість підвибірок відносно вихідної вибірки з єдиних позицій на основі принципів фрактального аналізу. Запропоновано методи визначення фрактальної розмірності вибірки, що оперують прямокутними блоками однакового розміру, покриваючи ними простір ознак: такий, що не враховує характеристики синтезованої моделі, такий, що враховує помилку (точність) синтезованої моделі, а також такий, що враховує точність і складність синтезованої моделі. Поряд із фрактальною розмірністю також запропоновано метод визначення показників якості вибірки на основі принципу масової розмірності стосовно до аналізу даних. Запропонований метод розбиває простір ознак на кластери однакового розміру і форми. Варіюючи розмір кластера, метод дозволяє одержувати різні рівні деталізації вибірки. Метод дозволяє визначити центр мас класу у вибірці, середню відстань між екземплярами кластера, нормоване середнє відхилення відстаней між екземплярами від їхнього середнього, масу і щільність екземплярів кластера, обсяг і площу поверхні прямокутного кластера, відношення обсягу до площі поверхні кластера, середньозважену рівномірність розташування екземплярів у кластерах класу, масу і щільність екземплярів класу, середньозважену рівномірність розташування екземплярів вибірки.

Результати. Розроблені показники реалізовані програмно і досліджені при вирішенні задачі класифікації ірисів Фішера.

Висновки. Проведені експерименти підтвердили працездатність запропонованого математичного забезпечення і дозволяють рекомендувати його для використання на практиці при вирішенні задач діагностування й автоматичної класифікації за ознаками. Перспективи подальших досліджень можуть полягати в створенні послідовних методів розрахунку комплексу запропонованих показників, оптимізації їхніх програмних реалізацій, а також експериментальному дослідженні запропонованих показників на більшому комплексі практичних задач різної природи і розмірності.

Ключові слова: вибірка, фрактальна розмірність, метрика якості, кластер, формування вибірок.

Subbotin S. A.
Dr. Sc., Professor, Head of the Department of Software Tools of Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine

THE FRACTAL DIMENSION BASED QUALITY METRICS OF DATA SAMPLES AND DEPENDENCE MODELS

Context. The problem of automating the sampling of the original sample a large amount for the construction of models precedent. The object of the study was to model quality samples to build the models precedents.

Objective. The goal of the work is the creation of a set of indicators to assess the quality of samples having a single nature, based on the principles of fractal analysis.

Method. A set of indicators is proposed to characterize the quality of the subsample with respect to the original sample with one point of view on the basis of the principles of fractal analysis. The methods of sample fractal dimension evaluation are proposed. They operating with rectangular blocks of equal size and covering by them the feature space. They are method not taking into account the characteristics of the synthesized model, method taking into account the error (accuracy) of synthesized model and method taking into account accuracy and complexity of the synthesized model. Along with the fractal dimension it is also provided a method for determining the sample quality indicators based on the principle of mass dimension with regard to data analysis. The proposed method divides the feature space on clusters of the same size and shape. The method allows obtaining different levels of sampling detail varying the size of the cluster. The method allows to determine the masses of the class center in the sample, the average distance between instances of the cluster, the normalized mean deviation of the distance between instances of their average mass and density of the instances of the cluster, the volume and surface area of rectangular cluster ratio of volume to surface area of the cluster, the weighted average of evenness of instances location in the clusters of a class, the mass and density of instances of the class, the weighted average of sample instances location.

Results. The developed indicators have been implemented in software and investigated for solving the problems of Fisher's Iris classification.

Conclusions. The conducted experiments have confirmed the proposed software operability and allow recommending it for use in practice for solving the problems of diagnosis and automatic classification on the features. The prospects for further research may include the creation of parallel methods for calculation of set of proposed indicators, the optimization of their software implementations, as well as a experimental study of proposed indicators on more complex practical problems of different nature and dimensionality.

Keywords: sample, fractal dimension, quality metric, cluster, sample formation.

REFERENCES

1. Jensen R., Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches. Hoboken, John Wiley & Sons, 2008, 339 p.
2. Chaudhuri A., Stenger H. Survey sampling theory and methods. New York, Chapman & Hall, 2005, 416 p. DOI: 10.1201/9781420028638
3. Ed. P. J. Lavrakas. Encyclopedia of survey research methods. Thousand Oaks, Sage Publications, 2008, Vol. 1–2, 968 p. DOI: 10.4135/9781412963947.n159
4. Subbotin S. A. Formirovaniye vyborok i analiz kachestva modeley na osnove neyronnykh i neyro-nechotkikh setey v zadachakh diagnostiki i raspoznavaniya obrazov: monografiya. Saarbrьcken, LAP Lambert academic publishing, 2012, 232 p. (ISBN 978-3-8473-4471-1).
5. Kokren U., per. s angl. Sonina I. M.; pod red. Volkova A. G., Druzhinina N. K. Metody vyborochnogo issledovaniya. Moscow, Statistika, 1976, 440 p.
6. Subbotin S. A. The training set quality measures for neural network learning, *Optical Memory and Neural Networks (Information Optics)*, 2010, Vol. 19, No. 2, pp. 126–139. DOI: 10.3103/s1060992x10020037
7. Subbotin S. A. Kompleks kharakteristik i kriteriyev sravneniya obuchayushchikh vyborok dlya resheniya zadach diagnostiki i raspoznavaniya obrazov, *Matematychni mashyny i systemy*, 2010, No. 1, pp. 25–39.
8. Subbotin S. A. Kriterii individual'noy informativnosti i metody otbora ekzemplyarov dlya postroyeniya diagnosticheskikh i raspoznayushchikh modeley, *Bionika intelektu,* 2010, No. 1, pp. 38–42.
9. Subbotin S. A. Metody formirovaniya vyborok dlya postroyeniya diagnosticheskikh modeley po pretsedentam, Visnyk Natsional'noho tekhnichnoho universytetu «Kharkivs'kyy politekhnichnyy instytut» : zb. nauk. prats. Kharkiv: NTU «KHPI», 2011, No. 17, pp. 149–156.
10. Roberts A., Cronin A. Unbiased estimation of multi-fractal dimensions of finite data sets, *Physica A: Statistical Mechanics and its Application,* 1996, Vol. 233, No. 3–4, pp. 867–878. DOI: 10.1016/s0378-4371(96)00165-3
11. Dubuc B., Quiniou J., Roques-Carmes C., Tricot C., Zucker S. Evaluating the fractal dimension of profiles, *Physical Review*, 1989, Vol. 39, No. 3, pp. 1500–1512. DOI:10.1103/PhysRevA.39.1500
12. Cheng Q. Multifractal Modeling and Lacunarity Analysis, *Mathematical Geology*, 1997, Vol. 29, No. 7, pp. 919–932. DOI:10.1023/A:1022355723781
13. Eftekhari A. Fractal Dimension of Electrochemical Reactions, *Journal of the Electrochemical Society*, 2004, Vol. 151, No. 9, pp. E291–E296. DOI:10.1149/1.1773583.
14. Popescu D. P., Flueraru C., Mao Y., Chang S., Sowa M. G. Signal attenuation and box-counting fractal analysis of optical coherence tomography images of arterial tissue, *Biomedical Optics Express*, 2010, Vol. 1, No. 1, pp. 268–277. DOI:10.1364/boe.1.000268
15. Li J., Du Q., Sun C. An improved box-counting method for image fractal dimension estimation, *Pattern Recognition*, 2009, Vol. 42, No. 11, pp. 2460–2469. DOI:10.1016/j.patcog.2009.03.001.
16. Crişan D. A., Dobrescu R. Fractal dimension spectrum as an indicator for training neural networks, *Universitatea Politehnica Bucuresti Sci. Bull. Series C*, 2007, Vol. 69, № 1, pp. 23–32.
17. Camastra F. Data Dimensionality Estimation Methods: A survey, *Pattern Recognition*, 2003, Vol. 36, Issue 12, pp. 2945–2954. DOI: 10.1016/S0031-3203(03)00176-6
18. Takens F. eds.: Braaksma B., Broer H. W., Takens F. On the numerical determination of the dimension of an attractor, *Dynamical Systems and Bifurcations*, Workshop, Groningen, 16–20 April 1984 : proceedings. Berlin, Springer, 1985, pp. 99–106. (Lecture Notes in Mathematics , Vol. 1125). DOI: 10.1007/bfb0075637
19. Chumak O. V. Entropii i fraktaly v analize dannykh. Moscow-Izhevsk, NITS «Regulyarnaya i khaoticheskaya dinamika», Institut komp'yuternykh issledovaniy, 2011, 164 p.
20. Zong-Chang Y. Establishing structure for artificial neural networks based-on fractal, *Journal of Theoretical and Applied Information Technology*, 2013, Vol. 49, No. 1, pp. 342–347.
21. Fisher Iris dataset [Electronic resource]. Access mode: https://archive.ics.uci.edu/ml/datasets/Iris