

УДК 004.93



КРИТЕРИИ ИНДИВИДУАЛЬНОЙ ИНФОРМАТИВНОСТИ И МЕТОДЫ ОТБОРА ЭКЗЕМПЛЯРОВ ДЛЯ ПОСТРОЕНИЯ ДИАГНОСТИЧЕСКИХ И РАСПОЗНАЮЩИХ МОДЕЛЕЙ

С.А. Субботин

Запорожский национальный технический университет, г. Запорожье, Украина, subbotin@zntu.edu.ua

С целью автоматизации процесса построения диагностических и распознающих моделей предложен комплекс критериев индивидуальной и групповой информативности экземпляров, методы формирования выборки, позволяющие оценивать качество, а также автоматизировать процесс формирования репрезентативных выборок по отношению к исходной выборке.

РАСПОЗНАВАНИЕ ОБРАЗОВ, ДИАГНОСТИКА, ВЫБОРКА, ЭКЗЕМПЛЯР, МЕТОДЫ ФОРМИРОВАНИЯ ВЫБОРОК

Введение

В задачах неразрушающей диагностики и распознавания образов по признакам одной из главных проблем является сокращение размерности выборки данных для построения модели принятия решений на основе машинного обучения по прецедентам [1].

Наряду с использованием методов отбора информативных признаков [1, 2], ставших традиционным инструментом сокращения размерности, размерность данных также возможно сократить путем выделения наиболее значимых для решения задачи распознавания экземпляров.

Известные методы извлечения выборок [1, 3–5], как правило, не обеспечивают гарантии выделения наиболее важных для задачи экземпляров, а также приводят к формированию избыточных выборок, поскольку основаны на случайном покрытии множества возможных решений. Поэтому представляется актуальной разработка математических критериев и методов, позволяющих оценивать значимость экземпляров и управлять процессом их выбора.

Целью данной работы является создание комплекса критериев оценивания индивидуальной и групповой информативности экземпляров, а также методов их выбора, позволяющих автоматизировать процесс формирования выборок для построения диагностических и распознающих моделей.

1. Постановка задачи

Пусть задана исходная выборка $\langle X, Y \rangle$. Необходимо извлечь из нее обучающую выборку (подвыборку) $\langle x, y \rangle$, которая является репрезентативной по отношению к исходной выборке, то есть отображает все наиболее важные ее свойства.

Обозначим: x^s – s -ый экземпляр выборки; x_j^s – значение j -го признака s -го экземпляра; y^s – значение выходного признака; сопоставленное s -му экземпляру выборки; S – число экземпляров выборки; S^q – количество экземпляров выборки, принадлежащих к q -му классу; N – число входных признаков; K – число классов.

2. Критерии индивидуальной информативности экземпляров

Задача критериев индивидуальной информативности экземпляров заключается в том, чтобы для каждого экземпляра можно было оценить его ценность для процесса построения диагностической или распознающей модели. Очевидно, что наибольшую ценность будут иметь те экземпляры, которые расположены на внутренних и внешних границах классов, а также близко расположены к центрам классов и кластеров. Соответственно зададим критерии, отражающие данные соображения.

Критерий индивидуальной информативности s -го экземпляра относительно внутренней границы классов определим как:

$$I_{\hat{G}}^s = \frac{1}{S-1} \sum_{p=1}^S \left\{ e^{-\alpha_{y^s, y^p} \sum_{j=1}^N (x_j^s - x_j^p)^2} \mid s \neq p, y^s \neq y^p \right\}$$

$$\text{или } I_{\hat{G}'}^s = \frac{1}{S-1} \sum_{p=1}^S \left\{ e^{-\alpha_{y^s, y^p} \min_{j=1,2,\dots,N} \{x_j^s - x_j^p\}} \mid s \neq p, y^s \neq y^p \right\}$$

$$\text{или } I_{\hat{G}}^s = \max_{p=1,2,\dots,S} \left\{ e^{-\alpha_{y^s, y^p} \sum_{j=1}^N (x_j^s - x_j^p)^2} \mid s \neq p, y^s \neq y^p \right\}$$

$$\text{или } I_{\hat{G}'}^s = \max_{p=1,2,\dots,S} \left\{ e^{-\alpha_{y^s, y^p} \min_{j=1,2,\dots,N} \{x_j^s - x_j^p\}} \mid s \neq p, y^s \neq y^p \right\},$$

где

$$\alpha_{k,q} = \frac{1}{S^k S^q \sum_{s=1}^S \sum_{p=1}^S \left\{ \sum_{j=1}^N (x_j^s - x_j^p)^2 \mid (y^s = k, y^p = q) \vee (y^s = q, y^p = k) \right\}}.$$

Данный критерий будет принимать значения в диапазоне от нуля до единицы: его значение будет тем больше, чем ближе экземпляр расположен к границе между разными классами.

Критерий индивидуальной информативности s -го экземпляра относительно его удаленности определим:

– относительно границы класса как:

$$I_U^s = 1 - \min_{p=1,2,\dots,S} \left\{ e^{-\alpha \sum_{j=1}^N (x_j^s - x_j^p)^2} \mid s \neq p, y^s = y^p \right\}$$

$$\text{или } I_{U'}^s = 1 - \min_{p=1,2,\dots,S} \left\{ e^{-\alpha \sum_{j=1,2,\dots,N} \min \{ |x_j^s - x_j^p| \}} \mid s \neq p, y^s = y^p \right\}.$$

$$\text{где } \alpha_k = \frac{1}{\max_{\substack{s=1,2,\dots,S; \\ p=s+1,\dots,S}} \left\{ \sum_{j=1}^N (x_j^s - x_j^p)^2 \mid y^s = y^p \right\}}.$$

– относительно внешней границы выборки как:

$$I_{U''}^s = 1 - \min_{p=1,2,\dots,S} \left\{ e^{-\alpha \sum_{j=1}^N (x_j^s - x_j^p)^2} \mid s \neq p \right\}$$

$$\text{или } I_{U'''}^s = 1 - \min_{p=1,2,\dots,S} \left\{ e^{-\alpha \sum_{j=1,2,\dots,N} \min \{ |x_j^s - x_j^p| \}} \mid s \neq p \right\}.$$

$$\text{где } \alpha = \frac{1}{\max_{\substack{s=1,2,\dots,S; \\ p=s+1,\dots,S}} \left\{ \sum_{j=1}^N (x_j^s - x_j^p)^2 \right\}}.$$

Данный критерий будет принимать значения в диапазоне от нуля до единицы: его значение будет тем больше, чем дальше экземпляр расположен по отношению к остальным экземплярам класса или выборки, то есть критерий будет выявлять редкие нетипичные случаи, а также экземпляры, расположенные на внешних границах классов.

Критерий индивидуальной информативности s -го экземпляра относительно его близости к центру класса (кластера) определим как:

$$I_O^s = \frac{1}{S^{y^s} - 1} \sum_{p=1}^S \left\{ e^{-\sum_{j=1}^N (x_j^s - x_j^p)^2} \mid s \neq p, y^s = y^p \right\}$$

$$\text{или } I_{O'}^s = \frac{1}{S^{y^s} - 1} \sum_{p=1}^S \left\{ e^{-\sum_{j=1,2,\dots,N} \min \{ |x_j^s - x_j^p| \}} \mid s \neq p, y^s = y^p \right\}.$$

Данный критерий будет принимать значения в диапазоне от нуля до единицы: его значение будет тем больше, чем ближе экземпляр расположен к “остову” класса.

Интегральный показатель информативности s -го экземпляра определим как:

$$\hat{I}^s = \max(I_{\bar{G}}^s, I_{\hat{G}}^s, I_O^s, I_U^s)$$

$$\text{или } \hat{I}'^s = \max(I_{\bar{G}'}^s, I_{\hat{G}'}^s, I_{O'}^s, I_{U'}^s)$$

$$\text{или } I^s = \max(I_{\bar{G}}^s, I_{\hat{G}}^s, I_O^s, I_U^s, I_{\bar{G}'}^s, I_{\hat{G}'}^s, I_{O'}^s, I_{U'}^s).$$

Данный критерий будет принимать значения в диапазоне от нуля до единицы: чем больше будет его значение, тем значимее s -ый экземпляр для построения модели, поскольку он либо находится

на границе между классами, либо является уникальным наблюдением, либо находится на внешней границе класса, либо соответствует “остову” класса (близок к центру кластера).

2. Критерии групповой информативности экземпляров

Задача критериев групповой информативности экземпляров заключается в том, чтобы, с одной стороны, охарактеризовать качество выборки, а с другой стороны, определить ее важнейшие характеристики для сравнения с другими выборками для решения той же задачи.

Критерий групповой информативности экземпляров выборки относительно описания внутренней границы классов определим как:

$$I_{\bar{G}} = \frac{1}{S} \sum_{s=1}^S I_{\bar{G}}^s \quad \text{или} \quad I_{\bar{G}'} = \frac{1}{S} \sum_{s=1}^S I_{\bar{G}'}^s,$$

$$\text{или } I_{\hat{G}} = \frac{1}{S} \sum_{s=1}^S I_{\hat{G}}^s \quad \text{или} \quad I_{\hat{G}'} = \frac{1}{S} \sum_{s=1}^S I_{\hat{G}'}^s.$$

Данный критерий будет принимать значения в диапазоне от нуля до единицы: его значение будет тем больше, чем больше выборка содержит экземпляров, расположенных на границе между разными классами.

Критерий групповой информативности выборки относительно описания внешних границ классов определим как:

$$I_U = \frac{1}{S} \sum_{s=1}^S I_U^s \quad \text{или} \quad I_{U'} = \frac{1}{S} \sum_{s=1}^S I_{U'}^s.$$

Данный критерий будет принимать значения в диапазоне от нуля до единицы: его значение будет тем больше, чем больше выборка содержит экземпляров, расположенных на внешних границах классов.

Критерий групповой информативности экземпляров выборки относительно описания центров классов (кластеров) определим как:

$$I_O = \frac{1}{S} \sum_{s=1}^S I_O^s \quad \text{или} \quad I_{O'} = \frac{1}{S} \sum_{s=1}^S I_{O'}^s.$$

Данный критерий будет принимать значения в диапазоне от нуля до единицы: его значение будет тем больше, чем больше выборка содержит экземпляров, расположенных близко к “остовам” классов.

Интегральный показатель групповой информативности экземпляров выборки определим как:

$$\bar{I} = \left\{ \frac{K}{4S} (I_{\bar{G}} + I_{\hat{G}} + I_O + I_U) \mid S \geq K, K \geq 2 \right\}.$$

Данный критерий будет принимать значения в диапазоне от нуля до единицы: чем больше будет его значение, тем более ценной является выборка

для построения модели, поскольку она менее избыточна и содержит наиболее важные экземпляры для аппроксимации границ и центров классов.

3. Переборные методы формирования выборок

Для формирования подвыборки из исходной выборки предлагается использовать *модифицированный метод полного перебора* [1], который будет включать следующие этапы:

1. Инициализация задать исходную выборку $\langle X, Y \rangle$. Оценить индивидуальные информативности для каждого экземпляра исходной выборки. Рассчитать значение критерия качества исходной выборки \bar{I} .

2. Генерация: сгенерировать все возможные подвыборки $\langle x(k), y(k) \rangle$ исходной выборки $\langle X, Y \rangle$, где k – номер подвыборки, $x(k), y(k)$ – соответственно экземпляры k -ой выборки и сопоставленные им значения выходного признака.

3. Для каждой сгенерированной подвыборки рассчитать значение критерия качества $\bar{I}(k)$.

4. Среди сформированных подвыборок $\{\langle x(k), y(k) \rangle\}$ в качестве решения $\langle x, y \rangle$ выбрать ту подвыборку $\langle x(p), y(p) \rangle$, которая наилучшим образом соответствует заранее заданному критерию выбора решения.

Предлагается использовать один из следующих критериев выбора решения:

– критерий максимума качества формируемой выборки: $p = \arg \max_k \{ \bar{I}(k) \}$;

– критерий максимального соответствия качества формируемой выборки качеству исходной выборки: $p = \arg \min_k \{ | \bar{I}(k) - \bar{I} | \}$.

Достоинством данного метода является то, что он гарантированно находит лучшее из возможных решений. Недостатком метода является его чрезвычайно высокая вычислительная сложность. Поэтому практическое применение полного перебора возможно только для небольших исходных выборок.

Альтернативой методу полного перебора может служить *модифицированный метод сокращенного перебора* [1] с добавлением и удалением экземпляров. Данный метод может быть представлен как последовательность следующих этапов:

1. Инициализация задать исходную выборку $\langle X, Y \rangle$. Оценить индивидуальные информативности для каждого экземпляра исходной выборки. Рассчитать значение критерия качества исходной выборки \bar{I} . Установить: $k = 1$.

2. Генерация базовой подвыборки $\langle x(1), y(1) \rangle$ исходной выборки $\langle X, Y \rangle$: включить в выборку $\langle x(1), y(1) \rangle$ все экземпляры исходной выборки, для которых:

$$I^s \geq \frac{\alpha}{S} \sum_{s=1}^S I^s,$$

где α – некоторая константа; $0 < \alpha \leq 1$.

3. Добавление экземпляров. Оценить для подвыборки $\langle x(k), y(k) \rangle$ значение показателя качества $\bar{I}(k)$. Если $\bar{I}(k) \leq \bar{I}(k-1)$, то принять: $x(k) = x(k-1)$, $y(k) = y(k-1)$, перейти к этапу 4; в противном случае: если $\bar{I}(k)$ является приемлемым, то перейти к этапу 4, в противном случае: установить: $k = k+1$, $x(k) = x(k-1)$, $y(k) = y(k-1)$; если в исходной выборке имеются экземпляры, отсутствующие в подвыборке $\langle x(k), y(k) \rangle$, то выбрать из них экземпляр с максимальным значением I^s и включить его в подвыборку $\langle x(k), y(k) \rangle$, после чего перейти к этапу 3; если в исходной выборке не осталось экземпляров, отсутствующих в $\langle x(k), y(k) \rangle$, то перейти к этапу 4.

4. Удаление экземпляров. Оценить для подвыборки $\langle x, y \rangle$ значение показателя качества $\bar{I}(k)$. Если $\bar{I}(k)$ является приемлемым и подвыборка $\langle x(k), y(k) \rangle$ – непустая, то установить $k = k+1$, $x(k) = x(k-1)$, $y(k) = y(k-1)$, исключить из $\langle x(k), y(k) \rangle$ экземпляр с минимальным значением I^s , после чего перейти к этапу 4; в противном случае – принять: $x(k) = x(k-1)$, $y(k) = y(k-1)$, перейти к этапу 5.

5. Вернуть в качестве решения выборку $\langle x(k), y(k) \rangle$.

Достоинством данного метода является то, что он сокращает количество рассматриваемых решений за счет включения наиболее перспективных решений в базовую подвыборку, которая затем последовательно дополняется оставшимися экземплярами до тех пор, пока продолжается увеличение групповой информативности выборки, после чего из выборки последовательно удаляются избыточные экземпляры до тех пор, пока групповая информативность оказывается приемлемой.

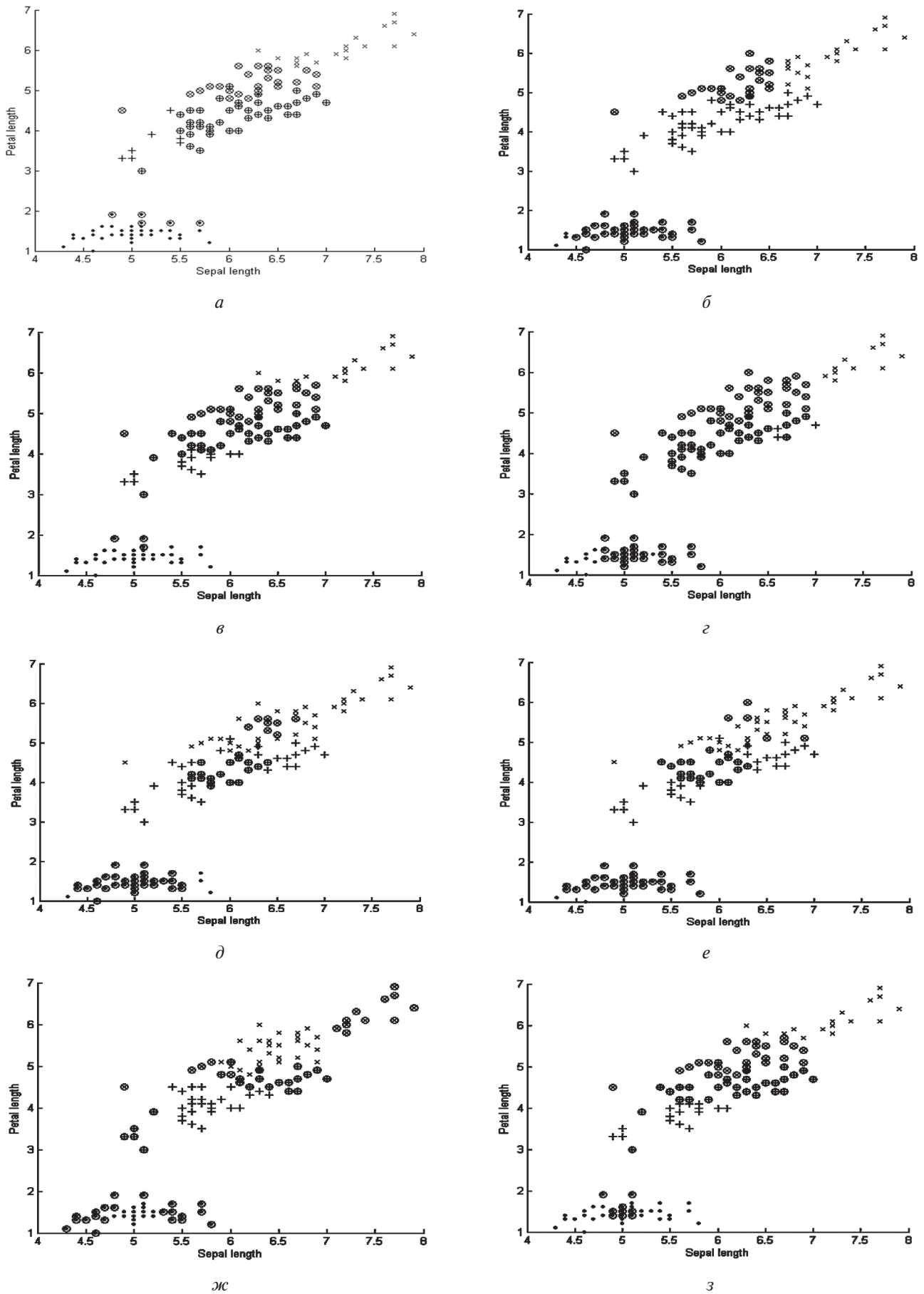
4. Эксперименты и анализ результатов

Для исследования практической применимости предложенных критериев и методов была разработана программа на языке пакета MATLAB, с помощью которой проводились эксперименты. Проведенные эксперименты подтвердили работоспособность предложенного математического обеспечения и его пригодность для решения практических задач диагностики и распознавания образов.

На рис. 1 изображена выборка для известной задачи распознавания ирисов Фишера в пространстве двух признаков (sepal length – длина чашелистика, petal length – длина лепестка).

Здесь значками “.” и “+” и “х”, обозначены, соответственно, экземпляры трех разных классов из исходной выборки, а значком “о” показаны экземпляры, имеющие индивидуальные оценки информативности выше средней по исходной выборке.

Легко видеть, что именно экземпляры, показанные значком “о”, расположены на границах и в центре классов, что наглядно демонстрирует работоспособность предложенных критериев.



ж

з

Рис. 1. Результаты выбора экземпляров на основе критериев:

$$a - I_G^s; \delta - I_G^s; \epsilon - I_G^s; \epsilon - I_G^s; \delta - I_O^s; e - I_O^s; \text{ж} - I_U^s; \text{з} - \hat{I}^s$$

Для сравнения методов формирования выборок оценим их временную сложность.

Для метода полного перебора сложность может быть оценена как $O((2^{S^*}-1)n)$, где S^* – количество экземпляров в исходной выборке, а n – сложность расчета интегрального показателя качества для подвыборки. Эта оценка свидетельствует о практической пригодности полного перебора только для исходных выборок небольшого объема.

Для модифицированного метода сокращенного перебора будем исходить из того, что базовая выборка, формируемая на втором этапе, будет включать не менее половины экземпляров исходной выборки; в свою очередь, на третьем этапе в худшем случае будет сгенерировано $0,5S^*$ подвыборок, а на четвертом этапе – S^* подвыборки. Следовательно, сложность метода можно оценить как $O((2+1,5S^*)n)$. Эта оценка свидетельствует о практической пригодности полного перебора только для исходных выборок большого объема. Таким образом, метод сокращенного перебора по сравнению с методом полного перебора будет работать быстрее в $\frac{2^{S^*}-1}{1,5S^*+2} \approx 2^{S^*-\log_2(1,5S^*)}$ раз.

Выводы

С целью автоматизации процесса построения диагностических и распознающих моделей в работе решена актуальная проблема анализа свойств и формирования обучающих выборок, репрезентативных относительно исходной выборки.

Научная новизна работы заключается в том, что: впервые предложен комплекс критериев индивидуальной и групповой информативности экземпляров выборки, позволяющий оценивать качество выборок, что позволяет автоматизировать процесс сравнения выборок между собой; получили дальнейшее развитие методы полного и сокращенного перебора, которые модифицированы путем введения разработанных критериев для оценивания пригодности экземпляров улучшать или ухудшать решения на основе индивидуальной и групповой информативностей, что позволяет автоматизировать процесс формирования репрезентативных выборок по отношению к исходной выборке.

Практическая ценность работы состоит в том, что: разработано программное обеспечение, реализующее предложенные критерии и методы; в результате проведенных экспериментов показана практическая пригодность разработанных критериев для отбора экземпляров; определены оценки сложности разработанных методов, позволяющие определить условия их применимости на практике.

Работа выполнена в рамках госбюджетной темы кафедры программных средств Запорожского национального технического университета “Информационные технологии автоматизации распознавания образов и принятия решений для диагностики в условиях неопределенности на основе гибридных нечеткологических, нейросетевых и мультиагентных методов вычислительного интеллекта” (номер государственной регистрации 0109U007673).

Список литературы: 1. Дубровин, В.И. *Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей* [Текст] : Монография / В. И. Дубровин, С. А. Субботин, А. В. Богуслаев, В. К. Яценко. – Запорожье: ОАО “Мотор–Сич”, 2003. – 279 с. 2. Субботин, С. О. *Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей* [Текст] : Монографія / С. О. Субботін, А. О. Олійник, О. О. Олійник ; під заг.ред. С. О. Субботіна. – Запоріжжя: ЗНТУ, 2009. – 375 с. 3. Джессен, Р. Дж. *Методы статистических обследований* [Текст] / Р.Д. Джессен; пер. с англ. Ю. П. Лукашина, Я. Ш. Паппэ; под ред. Е.М. Четыркина. – М.: Финансы и статистика, 1985. – 478 с. 4. Bernard, H. R. *Social research methods: qualitative and quantitative approaches* [Text] / H. R. Bernard. – Thousand Oaks: Sage Publications, 2006. – 784 p. 5. Кокрен, У. *Методы выборочного исследования* [Текст] / У. Кокрен; пер. с англ. И. М. Соннна; под ред. А. Г. Волкова, Н. К. Дружинина. – М.: Статистика, 1976. – 440 с.

Поступила в редколлегию 26.02.2010 г.

УДК 004.93

Критерії індивідуальної інформативності та методи відбору екземплярів для побудови діагностичних і розпізнавальних моделей / С.А. Суботін // Біоніка інтелекту: наук.-техн. журнал. – 2010. – № 1 (72). – С. 38–42.

З метою автоматизації процесу побудови діагностичних і розпізнавальних моделей запропоновано комплекс критеріїв індивідуальної та групової інформативності екземплярів, а також методи формування вибірки, що дозволяють оцінювати якість, а також автоматизувати процес формування репрезентативних вибірок відносно до вихідної вибірки.

Л. 1. Бібліогр.: 5 найм.

UDC 004.93

Criteria for individual informativity and sampling methods for diagnostic and recognition model building / S.A. Subbotin // Bionics of Intelligence: Sci. Mag. – 2010. – № 1 (72). – P. 38–42.

The set of criteria of individual and group informativity of exemplars, and sampling methods are proposed with the aim to automate the process of diagnostic and recognition model building. It allows to measure quality, and to automate the process of a sample formation representative to the original sample.

Fig. 1. Ref.: 5 items.