УДК 004.93'11

В.В. Искра, Н.А. Искра, М.М. Татур
Белорусский государственный университет информатики и радиоэлектроники
Республика Беларусь, 220072, г. Минск, ул. П. Бровки, 6

Влияние статистических характеристик обучающей выборки на её репрезентативность

V.V. Iskra, N.A. Iskra, M.M. Tatur
Belarusian State University of Informatics and Radioelectronics
Republic Belarus, 220072, Minsk, P. Brovka, 6

The Influence of the Statistical Characteristics of the Training Sample on its Representativeness

В.В. Іскра, Н.О. Іскра, М.М. Татур
Білоруський державний університет інформатики і радіоелектроніки
Республіка Білорусь, 220072, м. Мінськ, вул. П. Бровки, 6

Вплив статистичних характеристик навчальної вибірки на її репрезентативність

В статье рассматривается задача оценки репрезентативности выборки для обучения классификатора. Анализируется влияние статистических характеристик выборки на качество обучения. Предлагается определение репрезентативности через понятие функционала риска в рамках теории статистического обучения и проводится оценка состоятельности данного определения.
Ключевые слова: классификация, обучающая выборка, теория статистического обучения, репрезентативность.

The article considers the problem of estimating the representativeness of the sample for classifier training. The effect of statistical characteristics of the sample on the quality of training is analyzed. A definition of the concept of representativeness across functional risk in the framework of statistical learning is proposed and the consistency of this definition is evaluated.
Key words: classification, training sample, statistical learning theory, representativeness.

У статті розглядається задача оцінки репрезентативності вибірки для навчання класифікатора. Аналізується вплив статистичних характеристик вибірки на якість навчання. Пропонується визначення репрезентативності через поняття функціоналу ризику в рамках теорії статистичного навчання і проводиться оцінка спроможності даного визначення.
Ключові слова: класифікація, навчальна вибірка, теорія статистичного навчання, репрезентативність.

Введение

При практической оценке алгоритмов классификации важно отличать ошибки, вызванные несоответствиями выбранных алгоритмов, от ошибок, связанных с недостаточной репрезентативностью обучающей выборки. На данный момент не существует общепринятого математического определения понятия репрезентативности.

Целью данной работы является разработка подхода к определению понятия репрезентативности и исследование влияния статистических характеристик обучающей выборки на её репрезентативность для обоснования состоятельности полученного определения.

Постановка задачи. В соответствии с принципом минимизации эмпирического риска задача обучения классификатора с учителем представляется как минимизация функции, называемой функционалом риска. Ниже будет показано, что обучающая выборка является репрезентативной в той мере, в которой минимум соответствующего функционала эмпирического риска близок к минимуму функционала риска.

В целях оценки состоятельности данного определения понятия репрезентативности обучающей выборки, а также в целях определения его взаимосвязи с традиционными методами, основанными на вычислении вероятности смещения оценки математического ожидания значений параметров выборки при предположении их нормального распределения, экспериментально исследуются следующие величины:

– смещение среднего значения параметров обучающей выборки по сравнению с генеральной совокупностью;

– смещение дисперсии параметров обучающей выборки по сравнению с генеральной совокупностью;

– отклонение функционала эмпирического риска и минимума функционала риска для заданного классификатора;

– ошибка обобщения заданного классификатора.

Понятие репрезентативности

Понятие репрезентативности исследуемых данных используется в основном в социологии. Существует несколько подходов к определению данного термина:

Репрезентативную выборку можно определить как выборку, которая является (или считается) истинным отражением родительской популяции, то есть имеет тот же профиль признаков, например, возрастную структуру, классовую структуру, уровень образования [1].

Репрезентативность также можно понимать как «свойство выборочной совокупности воспроизводить параметры и значимые элементы структуры совокупности генеральной» [2], т.е. в этой трактовке, в отличие от приведённой выше, внимание акцентируется на значимых элементах совокупности.

С другой стороны, для социологического исследования важна не репрезентативность выборки, а репрезентативность результатов опроса [3]. Репрезентативность результатов опроса – это ситуация, когда совпадает распределение ответов на отдельный вопрос в выборочной и генеральной совокупностях. Следует отметить, что требование отражать «все свойства» генеральной совокупности представляется избыточным и даже недостижимым. Таким образом, можно принять следующее общее определение.

Репрезентативность – это способность выборки представлять параметры генеральной совокупности, значимые с точки зрения задач исследования.

Репрезентативность в контексте теории распознавания

Понятие репрезентативности данных в задачах распознавания, идентификации и классификации исследовано не достаточно глубоко.

Рассмотрим модель обучения классификатора с учителем [4], представленную на рис. 1.

На рис. 1 приняты следующие обозначения: $X,\{x \in X\}$ – генеральная совокупность; $F_x(x)$ – распределение вероятности; $d = f(x)$ – желаемый отклик (учитель);

$y = F(x,w)$ – обучаемая машина ($\boldsymbol{x}$ – вход, y – выход, $\boldsymbol{w}$ – свободные параметры); $T = \{(x_i, d_i)\}_{i=1}^{N}$ – обучающая выборка.

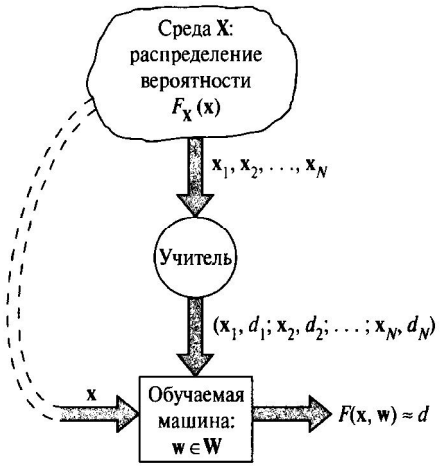

Рисунок 1 – Модель обучения классификатора с учителем

Задачу обучения, таким образом, можно рассмотреть как задачу аппроксимации, состоящую в нахождении функции $F(x,w)$, которая наилучшим образом приближает желаемую функцию $f(x)$.

Пусть $L(d,F(x,w))$ – мера несходства между желаемым откликом $d = f(x)$, соответствующим входному вектору x, и реальным откликом машины – $F(x,w)$ (в качестве L чаще всего используется квадрат Евклидовой нормы).

В общем виде реальное качество обучения можно представить так называемым функционалом риска:

$$R(w) = \int L(d,F(x,w))dF_{X,D}(x,d). \quad (1)$$

Целью обучения является поиск вектора $\boldsymbol{w}$, минимизирующего $R(\boldsymbol{w})$. Однако обобщённая функция распределения $F_{X,D}(x,d)$ в целом неизвестна. Единственный источник информации о ней – обучающая выборка T, которую можно использовать следующим образом:

Функционал $R(\boldsymbol{w})$ заменяется на функционал:

$$R_{emp}(w) = \sum_{i=1}^{N} L(d_i, F(x_i, w))\frac{1}{N}. \quad (2)$$

Формула (2) – так называемый функционал эмпирического риска [5].

На следующем шагу ищется $\boldsymbol{w} = argmin(R_{emp}(\boldsymbol{w}))$. При этом необходимым и достаточным условием состоятельности такой замены при достаточно большом значении количества примеров N является равномерная сходимость $R_{emp}(\boldsymbol{w})$ к $R(\boldsymbol{w})$ по вероятности при стремлении N к бесконечности [5], т.е.:

$$P\left(\sup_{w \in W} |R(w) - R_{emp}(w)| > \varepsilon\right) \to 0, \text{ при } N \to \infty. \quad (3)$$

Приведённый выше подход носит название «Принцип минимизации эмпирического риска» [4], [5]. В этом контексте понятие репрезентативности обучающей выборки можно сформулировать следующим образом.

Выборка T, обучающая машину $F(x,w)$ *с использованием функции стоимости* $L(d,F(x,w))$, *является репрезентативной в той мере, в которой минимум соответствующего функционала эмпирического риска* $R_{emp}(w)$ *близок к минимуму функционала риска* $R(w)$.

Можно принять и более сильное требование – близость функций $R_{emp}(w)$ и $R(w)$, однако на практике обычно вектор w вычисляется как аргумент, при котором градиент $R_{emp}(w)$ приблизительно равен нулю, т.е. значение имеют только экстремальные точки функций $R_{emp}(w)$ и $R(w)$.

Таким образом, можно ввести понятие *функционала риска репрезентативности*:

$$R_{repr}(w) = Q(R(w), R_{emp}(w)). \quad (4)$$

где Q – некоторый оператор, сравнивающий функции $R_{emp}(w)$ и $R(w)$.

В наиболее пессимистичном случае в качестве $Q(R(w), R_{emp}(w))$ берётся выражение:

$$Q(R(w), R_{emp}(w)) = \sup_{w \in W} |R(w) - R_{emp}(w)|. \quad (5)$$

При оценке эффективности обучения используются понятия ошибки обучения и ошибки обобщения [6]. Под ошибкой обучения подразумевается отклонение отклика машины от желаемого результата на примерах, взятых из обучающей выборки, а под ошибкой обобщения – на примерах, не вошедших в обучающую выборку.

Следует отметить, что в то время как ошибка обучения отражает скорее адекватность выбранных алгоритмов поставленной задаче, *ошибка обобщения напрямую зависит от способности данных обучающей выборки представлять генеральную совокупность*, т.е. буквально от репрезентативности обучающей выборки.

В целях формализации и теоретического обоснования этого тезиса приведём следующие аналитические выкладки. Представим генеральную совокупность $G = \{(x, f(x))\}_{x \in X}$ в виде $G = T \cup (G \setminus T) = T \cup A$. Здесь $A = G \setminus T$ – множество примеров генеральной совокупности, не попавших в обучающую выборку. Тогда функционал риска (1) можно представить следующим образом:

$$R(w) = \int_A L(d,F(x,w)) dF_{X,D}(x,d) + \int_T L(d,F(x,w)) dF_{X,D}(x,d). \quad (6)$$

Следует отметить, что $\int_T L(d,F(x,w)) dF_{X,D}(x,d)$ – просто иная форма записи $R_{emp}(w)$ из (2).

Обозначим первое слагаемое из (6) $\int_A L(d,F(x,w)) dF_{X,D}(x,d) = R_{gene}(w)$, и перепишем (6) в виде:

$$R(w) = R_{gene}(w) + R_{emp}(w) \quad (7)$$

или

$$R_{gene}(w) = R(w) - R_{emp}(w). \quad (8)$$

Поскольку $R_{gene}(w)$ – интеграл функции стоимости по всем примерам, не вошедшим в обучающую выборку, эту величину можно рассматривать как генерализованное выражение *ошибки обобщения*. Учитывая (8), условие (3) можно переписать в виде:

$$P\left(\sup_{w \in W} |R_{gene}(w)| > \varepsilon\right) \to 0, \text{ при } N \to \infty. \quad (9)$$

Для исследования зависимости репрезентативности выборки от $Q(R(w), R_{emp}(w))$ был проведён следующий эксперимент. В качестве генеральной совокупности был выбран известный набор данных – Ирисы Фишера [7]. В качестве классификатора – двухслойный персептрон, обучаемый по методу обратного распространения ошибки [4]. Для обучения классификатора использовались выборки ирисов фиксированных размеров для каждого класса. В качестве Q была взята евклидова норма. Результат эксперимента представлен на рис. 2:

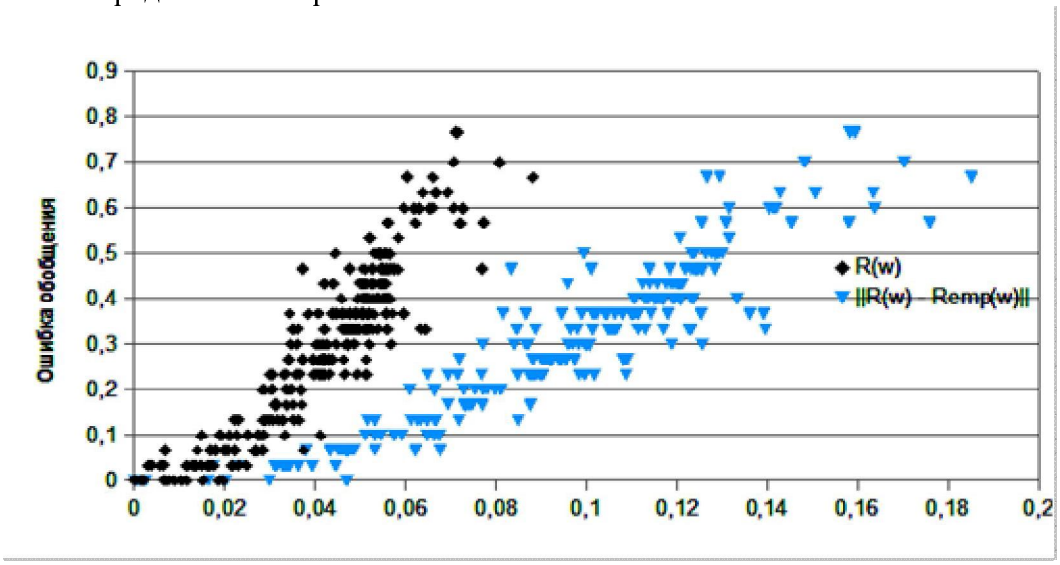

Рисунок 2 – Зависимость репрезентативности от функционала риска

На графике показано, что репрезентативность выборки действительно сильно коррелирует как с величиной $R(w)$, так и с $Q(R(w), R_{emp}(w))$.

Оценка репрезентативности

Существующие подходы априорного определения репрезентативности выборки из генеральной совокупности сводятся в основном к подсчёту вероятности смещения оценки математического ожидания интересующих параметров (т.е. компонентов вектора $\boldsymbol{x}$) при предположении о нормальном распределении самого вектора $\boldsymbol{x}$ [8]; а также к неформальной и нередко интуитивной оценке структуры выборки [3].

Для изучения влияния смещения статистических моментов (математического ожидания и дисперсии) в выборке по сравнению с генеральной совокупностью проведём эксперимент на Ирисах Фишера, подобный описанному выше.

Для обучения классификатора будем использовать выборки ирисов фиксированных размеров для каждого класса и оценим влияние отклонения среднего по выборке значения и стандартного отклонения параметров от математического ожидания и дисперсии в генеральной совокупности соответственно.

На рис. 3 и 4 показаны графики полученных зависимостей. Нетрудно заметить, что, по крайней мере, для выбранной исходной генеральной совокупности репрезентативность выборки практически не коррелирует с соответствием среднего значения и стандартного отклонения параметров выборки от соответствующих характеристик генеральной совокупности.

Для учёта структурных различий выборки T и генеральной совокупности G (например, количества и расположения кластеров и т.п.) можно воспользоваться сле-

дующим наблюдением. Проецирование векторов G на направления собственных векторов матрицы ковариации, соответствующих доминирующим собственным значениям, имеет тенденцию раскрывать кластерную структуру данных [4, с. 523]. Следовательно, можно ожидать, что искажение внутренней структуры данных в выборке T выразится в значительном изменении разложения матрицы ковариации T на собственные векторы по отношению к разложению матрицы ковариации G.

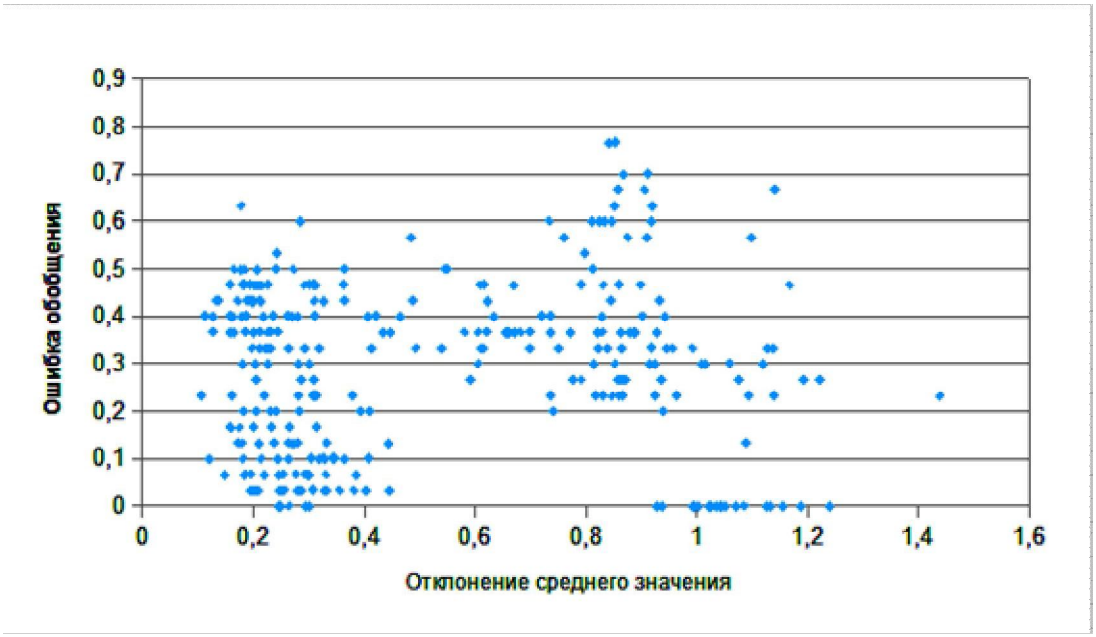

Рисунок 3 – Зависимость репрезентативности от отклонения среднего значения параметров в выборке по отношению к генеральной совокупности

Рисунок 4 – Зависимость репрезентативности от смещения стандартного отклонения параметров в выборке по отношению к генеральной совокупности

На рис. 5 показана зависимость репрезентативности от отклонения собственных векторов матрицы ковариации:

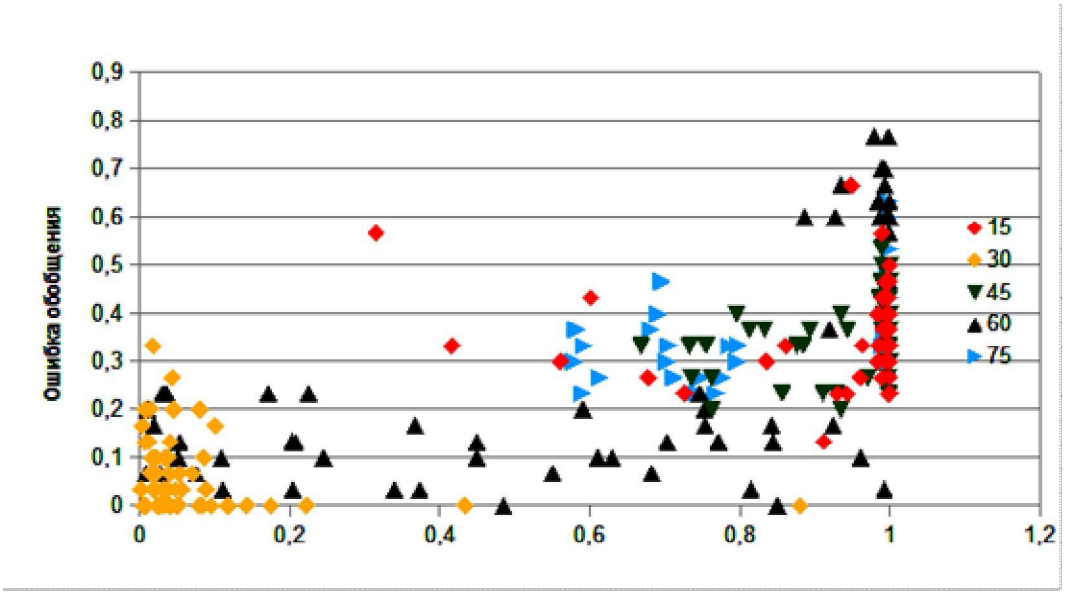

Рисунок 5 – Зависимость репрезентативности от отклонения собственных векторов матрицы ковариации

В качестве меры отклонения собственных векторов (ось абсцисс) используется величина:

$$r = 1 - \left(\prod_{i=1}^{M} \left(v_i^s, v_i^g \right) \right); v^s \in V^s, v^g \in V^g \ . \qquad (10)$$

где M – количество признаков, $v^s \in V^s, v^g \in V^g$ – множества собственных векторов матриц ковариации выборки и генеральной совокупности соответственно. Маркерам различного вида на графике соответствуют выборки разного размера.

Видно, что в граничных случаях, когда отклонение либо слишком мало (близко к нулю) или слишком велико (близко к единице), корреляция слаба, однако при средних значениях величины r наблюдается чёткая зависимость между максимальной ошибкой и r.

Выводы

Экспериментальная проверка показала, что использование тривиальных оценок репрезентативности, основанных на сравнении средних значений и отклонений параметров выборки с генеральной совокупностью, может давать неудовлетворительный результат для задач классификации.

Наиболее точным образом репрезентативность определяется через понятие функционала риска в рамках теории статистического обучения. Однако такой способ предполагает довольно глубокую информацию о генеральной совокупности и даёт результат, привязанный к конкретному типу классификатора.

Наиболее перспективным представляется подход, основанный на вычислении собственных векторов матрицы ковариации признаков. Этот подход позволяет, с одной стороны, определить репрезентативность обучающей выборки по отношению к генеральной совокупности, если известна ковариация признаков в генеральной совокупности, а с другой – выяснить, уменьшается ли репрезентативность обучающей выборки при её сокращении.

Важно отметить, что предложенная в данной статье метрика накладывает только нижнее ограничение на репрезентативность, для более конкретного результата необ-

ходимо глубже изучать взаимосвязь между репрезентативностью выборки и собственными векторами матрицы ковариации признаков.

Литература

1. Большой толковый социологический словарь (Collins) : в 2 т. , пер. с англ. – М. : Вече, АСТ, 1999. – Т. 2 (П-Я). – С. 158.
2. Сотникова Г.Н. Репрезентативность / Г.Н. Сотникова, Г.В. Осипов // Российская социологическая энциклопедия. – М. : НОРМА-ИНФА-М, 1998. – С. 445.
3. Ильясов Ф.Н. Репрезентативность результатов опроса в маркетинговом исследовании / Ф.Н. Ильясов // Социологические исследования. – 2011. – № 3. – С. 112-116.
4. Хайкин С. Нейронные сети : полный курс / Хайкин С. – М. : Издательский дом «Вильямс», 2006. – С. 140-146.
5. Vapnik V.N. Principles of risk minimization for learning theory / V.N. Vapnik // Advances in Neural Information Processing Systems. – 1992. – Vol. 4. – P. 831-838.
6. Vapnik V.N. Statistical Learning Theory / Vapnik V.N. – New York : Wiley, 1998.
7. Fisher R.A. The Use of Multiple Measurements in Taxonomic Problems / Fisher // R.A. Annals of Eugenics. – 1936 – 7. – P. 179-188.
8. Telhaj Sh. Representativeness of Yellis Sample Schools in a Study of Subject Enrollment of 14-16 olds / Sh. Telhaj, D. Hutton D. // Economics and Social Research Council, 2009.

Literature

1. Big Explanatory Dictionary of Sociology (Collins) // In 2 vol. – vol. 2. – Moscow. : Veche, ACT, 1999. – P. 158.
2. Sotnikova G.N. Representativeness / G.N. Sotnikova, G.V. Osipov // Russian sociological encyclopedia. – Moscow: NORMA-INFA-M, 1998. – P. 445.
3. Il'yasov F. N. Representativeness of the survey results in the marketing research // Sociological Studies, 2011. – № 3. – P. 112-116.
4. Haykin S. Neural Networks: A full course / S. Haykin. - Moscow: Publishing house "Williams", 2006. – P. 140-146.
5. Vapnik V. N. Principles of risk minimization for learning theory // Advances in Neural Information Processing Systems, 1992. – vol. 4. – P. 831-838.
6. Vapnik V.N. Statistical Learning Theory, New York : Wiley, 1998
7. Fisher R.A. The Use of Multiple Measurements in Taxonomic Problems / Annals of Eugenics 7, 1936 – P. 179-188.
8. Telhaj Sh. Representativeness of Yellis Sample Schools in a Study of Subject Enrollment of 14-16 olds / Sh. Telhaj, D. Hutton D. // Economics and Social Research Council, 2009.

REZUME

V.V. Iskra, N.A. Iskra, M.M. Tatur

The Influence of the Statistical Characteristics of the Training Sample on its Representativeness

This article examines the problem of estimating the representativeness of the sample for classifier training. A definition of the concept of representativeness across functional risk in the framework of statistical learning is proposed and the consistency of this definition is evaluated. An approach to the definition of representativeness as proximity sign of decomposition of the covariance matrix in the sample relative to the general population is also proposed. In order to assess the consistency of the definition of representative training sample, and to determine its relationship with the traditional methods based on the calculation of the probability of bias in the expectation values of the parameters of the sample under the assumption of a normal distribution, the following values are experimentally investigated:

– the shift of the average values of the parameters of training sample compared to the general population;

– the shift of the dispersion parameters of training sample compared to the general population;

– the deviation of the empirical risk functional and minimal risk functional for the given classifier.

Статья поступила в редакцию 16.04.2013.