*Информационные процессы, Том 18, № 2, 2018, стр. 118–124*


**ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ**

# Метод классификации данных и функции распределения

**В. А. Нестеренко**

*Институт математики, механики и компьютерных наук,*
*Южный федеральный университет, Ростов-на-Дону, Россия*
*neva09@mail.ru*

Поступила в редколлегию 30.05.2018

**Аннотация**—В статье предлагается новый метод классификации данных. Метод основан на сравнения функций распределения тестовой и обучающей выборок. По результатам сравнения делается вывод о принадлежности объектов тестовой выборки классу, представленному обучающей выборкой. Сформулирован критерий принадлежности объекта искомому классу. Приведены примеры применения предлагаемого метода. Данный метод может быть использован при классификации данных в задачах Data Mining.

**КЛЮЧЕВЫЕ СЛОВА:** классификация, классификация данных, функция распределения, статистика хи-квадрат.

## 1. ВВЕДЕНИЕ

В общем случае задача классификации данных заключается в том, чтобы множество объектов отобразить на конечное множество классов [1]. Правила отображения формируются путём выявления степени сходства рассматриваемого объекта с объектами класса. Для этого могут быть использованы разные критерии [2,3]: расстояние в пространстве характеристик, вероятность вхождения объекта в обучающую выборку, применение различных классификаторов и так далее. В любом случае, существующие методы классификации используют сопоставление свойств единичного объекта со свойствами группы объектов - класса. В данной статье предлагается новый метод классификации [4], основанный на использовании отношения свойств группы объектов со свойствами класса. По результатам сравнения функций распределения тестовой и обучающей выборок делается вывод о возможности вхождения объектов тестовой выборки в класс. Для этого используется степень сходства функции распределения объектов тестовой выборки и функции распределения искомого класса. Предлагаемый метод не требует введения метрики в пространстве состояний, он основан на использовании статистики критерия согласия $\chi^2$.

## 2. ИДЕЯ МЕТОДА КЛАССИФИКАЦИИ

Рассмотрим множество объектов $\mathbf{U} = \{U_1, ..., U_N\}$. Пусть каждому объекту $U_k \in \mathbf{U}$ соответствуют $D$ характеристик. Набор объектов $\mathbf{U}$ можно представить точками в пространстве размерности $D$, координатные оси пространства соответствуют характеристикам объектов. Назовём это пространство пространством признаков или характеристик.

Пусть множество $\mathbf{U}$ состоит из объединения множеств $\mathbf{A}$ и $\mathbf{Z}$. В этом случае классификация данных будет заключаться в выделении из множества $\mathbf{U}$ объектов, принадлежащих множеству $\mathbf{A}$. Обозначим через $\mathbf{E} = \{E_1, ..., E_{n_E}\}$ обучающую выборку из множества $\mathbf{A}$ и через $\mathbf{T} = \{T_1, ..., T_{n_T}\}$ - случайную выборку из множества $\mathbf{U}$. Определить принадлежность элементов тестовой выборки $\mathbf{T}$ классу $\mathbf{A}$ можно при помощи проверки критерия однородности для тестовой $\mathbf{T}$ и обучающей $\mathbf{E}$ выборок. Если гипотеза об однородности принимается, то обучающая

и тестовая выборки имеют одинаковые функции распределения и объекты тестовой выборки принадлежат тому же классу, что и объекты обучающей выборки. Такое прямолинейное использование критерия согласия имеет существенный недостаток. Так как вероятность того, что все объекты случайной тестовой выборки принадлежат искомому классу мала $\sim (N_A/N)^{n_T}$, то необходимо перебрать много тестовых выборок для поиска объектов из класса **A**.

Что бы избежать указанного недостатка, используем две тестовые выборки одновременно и выделим ту из них, которая содержит больше элементов искомого класса **A**. Будем считать, что множества **A** и **Z** состоят из $N_A$ и $N_Z$ точек соответственно, $n$ - число точек в случайной тестовой выборке. Пусть $P(i, n-i, N_A, N_Z)$ вероятность получить в случайной выборке $i$ точек из множества **A** и $n-i$ - из множества **Z**. Для двух последовательных выборок без возврата вероятность получить значения $(i, n-i)$ и $(j, n-j)$ для элементов из множеств **A** и **Z** составит

$$P(i, n-i, N_A, N_Z)\, P(j, n-j, N_A-i, N_Z-n+i)$$

Упорядочим эти тестовые выборки так, чтобы $i \geq j$, т.е. число объектов из множества **A** во второй выборке было меньше чем в первой. Если объект из **A** попал в одну из двух тестовых выборок, то вероятность его попадания в первую равна $P_A(i,j) = i/(i+j)$. Так как $i \geq j$, то $P_A(i,j) \geq 0.5$. Для сравнения тестовых выборок с обучающей можно использовать критерий согласия функций распределения, в этом случае вероятность $P_A(i,j)$ будет выше для той выборки, для которой критерий согласия даёт лучший результат. Таким образом, при упорядочивании выборок по числу элементов из **A** можно использовать статистику критерия согласия.

Ассоциируем с объектами $U_k \in \mathbf{U}$ счётчики $c_k^{(+)}$ и $c_k^{(-)}$. В каждом испытании будем увеличивать значение счётчиков $c_k^{(+)}$ на 1 для объектов первой выборки и увеличивать значения счётчиков $c_k^{(-)}$ для объектов второй. Средние по выборкам значения счётчиков $\overline{c}_k^{(+)}$ и $\overline{c}_k^{(-)}$ для $U_k \in \mathbf{A}$ удовлетворяют соотношению:

$$\frac{\overline{c}_k^{(+)}}{\overline{c}_k^{(+)} + \overline{c}_k^{(-)}} = \frac{\sum\limits_{i=0}^{n} P(i, n-i, N_A, N_Z) \sum\limits_{j=0}^{i} \dfrac{i}{i+j} P(j, n-j, N_A-i, N_Z-n+i)}{\sum\limits_{i=0}^{n} P(i, n-i, N_A, N_Z) \sum\limits_{j=0}^{i} P(j, n-j, N_A-i, N_Z-n+i)} \tag{1}$$

Так как выборки упорядочены и в формуле (1) суммирование производится по области значений $i \geq j$, то $\overline{c}_k^{(+)} / \left(\overline{c}_k^{(+)} + \overline{c}_k^{(-)}\right) \geq 0.5$ если $U_k \in \mathbf{A}$.

Аналогичные рассуждения для объектов множества **Z** дают:

$$\frac{\overline{c}_k^{(+)}}{\overline{c}_k^{(+)} + \overline{c}_k^{(-)}} = \frac{\sum\limits_{i=0}^{n} P(i, n-i, N_A, N_Z) \sum\limits_{j=0}^{i} \dfrac{n-i}{2n-i-j} P(j, n-j, N_A-i, N_Z-n+i)}{\sum\limits_{i=0}^{n} P(i, n-i, N_A, N_Z) \sum\limits_{j=0}^{i} P(j, n-j, N_A-i, N_Z-n+i)} \tag{2}$$

и $\overline{c}_k^{(+)} / \left(\overline{c}_k^{(+)} + \overline{c}_k^{(-)}\right) < 0.5$ если $U_k \in \mathbf{Z}$.

Таким образом, сравнивая две случайные выборки и изменяя значения счётчиков $c_k^{(+)}$ и $c_k^{(-)}$ соответствующим образом, после достаточного числа испытаний, в соответствии с формулами (1) и (2), мы приходим к следующему результату:

$$\begin{aligned} p_k &\geq 0.5 \text{ если } U_k \in \mathbf{A} \\ p_k &< 0.5 \text{ если } U_k \notin \mathbf{A} \end{aligned} \tag{3}$$

где

$$p_k = \frac{c_k^{(+)}}{c_k^{(+)} + c_k^{(-)}} \tag{4}$$

В предлагаемом методе соотношения (3) используются в качестве критерия классификации объектов при определении их принадлежности искомому классу.

## 3. МЕРА БЛИЗОСТИ ФУНКЦИЙ РАСПРЕДЕЛЕНИЯ

При выводе критерия классификации (3) основным моментом было выделение из двух тестовых выборок той, которая содержит больше элементов из класса **A**. Как отмечалось ранее, для нахождения такой выборки можно использовать статистику какого-либо критерия согласия функций распределения. При этом выполнение самого критерия согласия не требуется, достаточно из двух тестовых выборок выбрать ту, которая имеет "лучшее согласие" с обучающей выборкой. В рассматриваемой статье для этих целей используется статистика критерия согласия $\chi^2$ Пирсона [5]. В отличии от других, статистика этого критерия легко обобщается на многомерный случай.

Пусть каждому объекту $U_k \in \mathbf{U}$ соответствует набор признаков $X_k^{(i)}$ $1 \le i \le D$, здесь $D$ - размерность пространства характеристик. Для выравнивания масштабов характеристик по разным осям нормируем характеристики обучающей выборки $\mathbf{E} \subset \mathbf{A}$ на интервал $[-1, +1]$ и масштабируем характеристики всего набора данных **U**:

$$\widetilde{X}_k^{(i)} = \frac{\left(2\,X_k^{(i)} - X_{max}^{(i)} - X_{min}^{(i)}\right)}{\left(X_{max}^{(i)} - X_{min}^{(i)}\right)} \quad \text{где} \quad X_{max}^{(i)} = \max_{\{k:U_k \in E\}} X_k^{(i)} \quad \text{и} \quad X_{min}^{(i)} = \min_{\{k:U_k \in E\}} X_k^{(i)}$$

Так как статистика $\chi^2$ предполагает группирование элементов выборки, то разделим пространство характеристик на $B$ непересекающихся частей так, чтобы точки обучающей выборки **E** равномерно распределились среди этих частей пространства:

$$n_i^{(E)} \approx n_j^{(E)} \quad 1 \le i, j \le B$$

(способ разбиения приведён в ПРИЛОЖЕНИИ).

Если $n_E$ обозначает размер обучающей выборки, $n_T$ - размер тестовой выборки, $n_i^{(E)}$ и $n_i^{(T)}$ - частоты попадания элементов соответствующих выборок в часть пространства с номером $i$, то статистика двухвыборочного критерия $\chi^2$ имеет вид:

$$\chi^2_{n_E,n_T} = \sum_{i=1}^{B} \frac{1}{\left(n_i^{(E)} + n_i^{(T)}\right)} \cdot \left(\frac{n_i^{(E)}}{n_E} - \frac{n_i^{(T)}}{n_T}\right)^2 \tag{5}$$

Теперь, при сравнении двух тестовых выборок, "лучшей" (содержащей больше элементов из множества **A**) мы будем считать ту, для которой статистика (5) принимает меньшее значение.

Принятый способ разбиения (результат показан на Рис7) приводит к появлению выделенных направлений в пространстве характеристик, связанное с выбранным направлением координатных осей. Это нарушает симметрию исходной задачи и ухудшает качество классификации. Для восстановления равноправия разных направлений в пространстве поступим следующим образом: в каждой координатной плоскости совершим s раз поворот относительно начала координат на угол $\pi/2(s+1)$, для каждого нового направления координатных осей

проведём новые испытания и изменим значения счётчиков $c_k^{(+)}$, $c_k^{(-)}$ соответствующим образом. В пространстве размерности $D > 1$ операции поворота следует выполнять независимо во всех $D(D-1)/2$ координатных плоскостях. Опыт показывает, что достаточно $s = 3$ для устранения эффекта выделенного направления.

## 4. ПРИМЕРЫ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ МЕТОДА

Для демонстрации примеров применения предлагаемого метода классификации используем различные конфигурации множеств **A** и **Z** в двумерном пространстве характеристик. Двумерный случай выбран из-за возможности наглядного представления результатов классификации. Рассматриваемый метод применим для произвольного числа характеристик объектов.

Во всех примерах множества **A** и **Z** содержат $N_A = 400$ и $N_Z = 600$ точек соответственно, обучающая и тестовые выборки состоят из $n = 100$ точек каждая. При использовании статистики $\chi^2$ (5) пространство характеристик разбивается на $B = 10$ областей. Для лучшей наглядности на рисунках отображена только каждая 3-я точка множества **U**. На всех рисунках красным цветом обозначены точки исходно принадлежащие множеству **A**, синим - множеству **Z**, зелёные треугольники обозначают элементы обучающей выборки **E**. Точки, отнесённые в результате классификации к классу **A**, обозначены кружками с сохранением цвета исходного множества. В подписях к рисункам приведены оценки качества классификации: $TPR$ - true positive rate, $FPR$ - false positive rate.

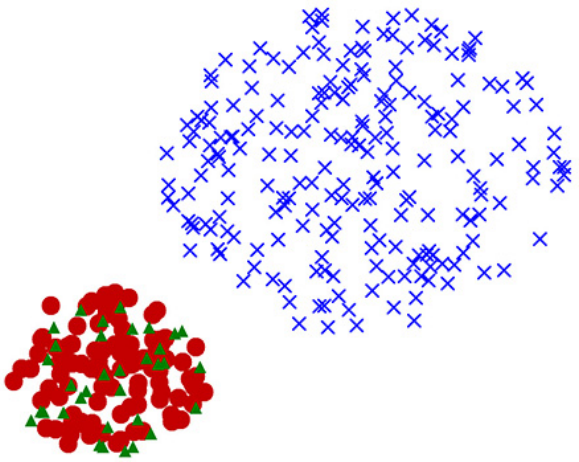

Рис 1. $TPR = 1.00$ $FPR = 0.00$

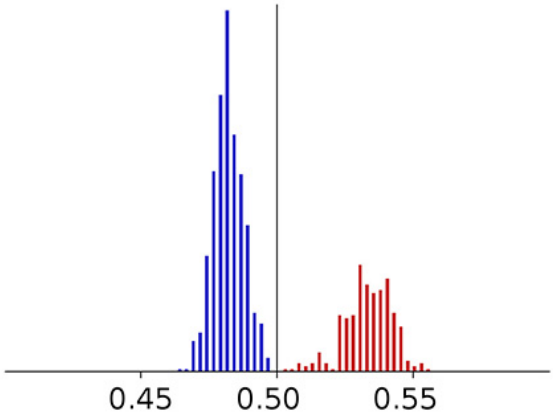


Рис 2.

На Рис1 приведены результаты для случая изолированных в пространстве множеств **A** и **Z**. На Рис2 представлена гистограмма распределения значений параметра $p_k$ (4) для точек множества **U**. Как и ранее, красный цвет соответствует объектам, изначально принадлежащим классу **A**, синий - не принадлежащим. Как видно из гистограммы на Рис2 результаты классификации приводят к разделению объектов по параметру $p_k$ относительно значения 0.5 в соответствии с критерием (3). Так как исходные данные представлены изолированными множествами, то и полученная гистограмма состоит из двух изолированных частей.

Для пересекающихся в пространстве множеств результаты классификации приведены на Рис3. В этом случае данные, соответствующие разным множествам, на гистограмме Рис4 не изолированы. Часть точек из множества **Z** находятся в области пространства множества **A**. Тем не менее, как видно из результатов на гистограмме Рис4, и в этом случае параметр $p_k$ также может быть критерием классификации относительно значения 0.5.

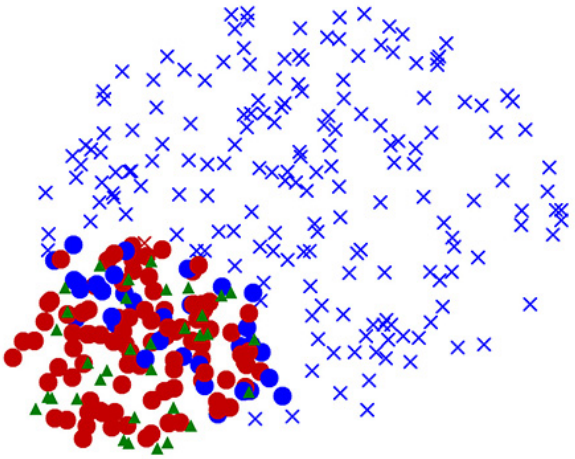

Рис 3. $TPR = 0.98$ $FPR = 0.15$

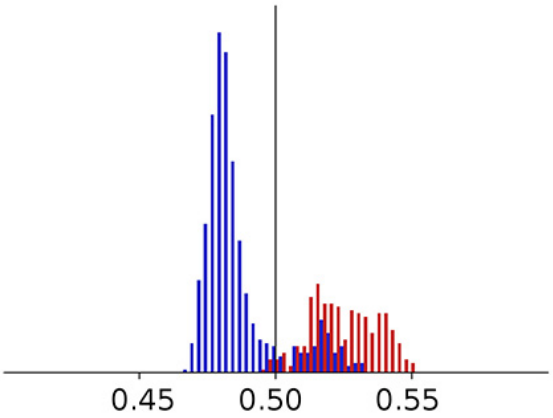


Рис 4.

Результат классификации для конфигурации множеств приведённой на Рис5 свидетельствует о том, что данный метод классификации не является линейным методом.

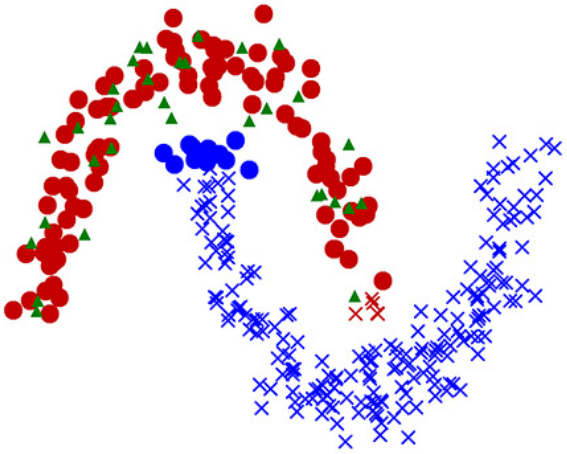

Рис 5. $TPR = 0.95$ $FPR = 0.05$

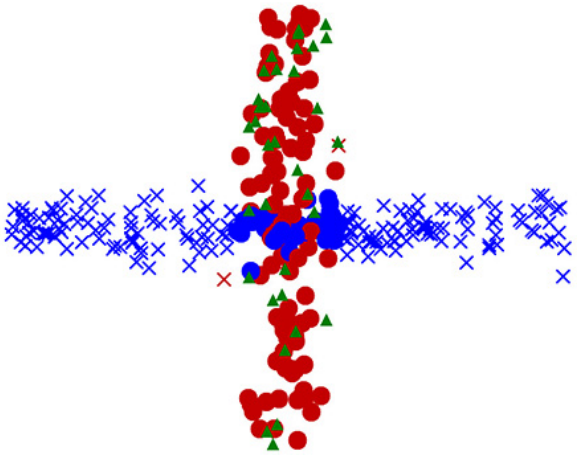

Рис 6. $TPR = 0.98$ $FPR = 0.21$

Увеличение числа ложных срабатываний при классификации в примерах на Рис3 и Рис6 вызвано тем, что точки множества **Z**, находящиеся в области пересечения, идентифицируются как принадлежащие классу **A** (синие кружки на Рис3, Рис6). Объекты, находящиеся на пересечении множеств **A** и **Z**, имеют одинаковые характеристики и неразличимы в рамках решаемой задачи.

## 5. ЗАКЛЮЧЕНИЕ

В статье рассмотрен новый метод классификации данных. Особенностью предлагаемого метода является использование при классификации совместных свойств группы объектов, а не свойств отдельных объектов по отношению к классу. Метод основан на применении статистики критерия согласия $\chi^2$ Пирсона, статистика используется для определения степени "близости" функций распределения обучающей и тестовой выборок. На основе сравнения функций распределения делается вывод о возможной принадлежности объектов искомому классу. Описание предлагаемого метода приведено для проверки объектов на принадлежность одному классу.

Метод легко может быть обобщён на случай с произвольным числом классов. Предлагаемый метод не требует введения метрики в пространстве характеристик. Для его использования достаточно упорядочить данные в соответствии со значениями характеристик. Метод может быть использован при классификации данных с числовыми и категориальными признаками.

В работе приведены результаты применения рассматриваемого метода к тестовым наборам данных. Показаны примеры хороших результатов классификации и случаи не столь хорошего качества классификации. Приведённые результаты позволяют сделать вывод о возможности применения предлагаемого метода в задачах Data Mining.

## *ПРИЛОЖЕНИЕ*

Использование статистики критерия $\chi^2$ предполагает разбиение пространства характеристик на $B$ непересекающиеся части так, чтобы точки обучающей выборки $\mathbf{E}$ равномерно распределились среди этих частей: $n_i^{(E)} \approx n_j^{(E)}$ $1 \le i, j \le B$, где $n_i^{(E)}$ - частота попадания элементов обучающей выборки в часть пространства с номером $i$.

Пусть $D$ - размерность пространства и $x_{min}^{(i)}$, $x_{max}^{(i)}$ $1 \le i \le D$ - минимальные и максимальные значения соответствующих координат точек множества $\mathbf{U}$:

$$x_{min}^{(i)} = \min_{\{k:U_k \in \mathbf{U}\}} \tilde{X}_k^{(i)} \ , \quad x_{max}^{(i)} = \max_{\{k:U_k \in \mathbf{U}\}} \tilde{X}_k^{(i)}$$

На каждой итерации алгоритма разбиению подлежит область пространства $x_L^{(i)} \le x^{(i)} \le x_R^{(i)}$ $1 \le i \le D$.

*Шаг 1.* Устанавливаем начальные значения границ области разбиения:

$$x_L^{(i)} = x_{min}^{(i)} \ , \quad x_R^{(i)} = x_{max}^{(i)} \quad 1 \le i \le D$$

*Шаг 2.* Выбираем координату $m$, вдоль которой область разбиения имеет наибольший размер $x_R^{(m)} - x_L^{(m)}$.

*Шаг 3.* Начиная от значения $x = x_L^{(m)}$ увеличиваем $x$ до те х пор, пока в части пространства $\{x_L^{(m)} \le x^{(m)} < x,\ x_L^{(i)} \le x^{(i)} \le x_R^{(i)}$ при $i \ne m\}$ не окажется приблизительно $n_E/B$ точек, $n_E$ - размер обучающей выборки. Отмечаем полученную часть как одну из искомых частей разбиения.

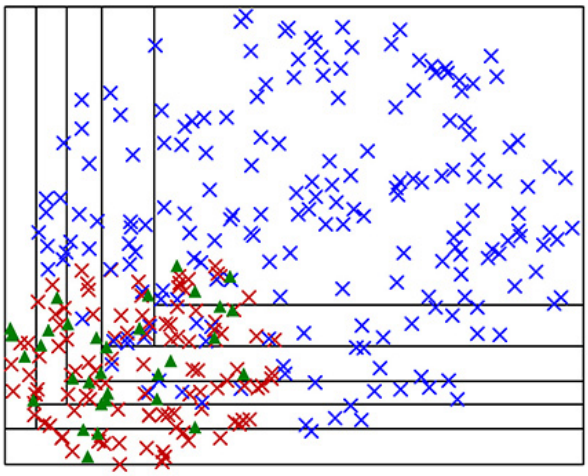

Рис 7.

*Шаг 4.* Устанавливаем новое граничное значение $x_L^{(m)} = x$ и возвращаемся к шагу 2. Последовательность шагов 2-4 повторяем $B-1$ раз и в результате получаем $B-1$ частей пространства. Оставшаяся область $x_L^{(i)} \le x^{(i)} \le x_R^{(i)}$, с полученными на предыдущих шагах граничными значениями $x_L^{(i)}$ и $x_R^{(i)}$, является последней частью разбиения пространства. Таким образом, в результате описанного процесса получаем $B$ частей пространства, с примерно одинаковым числом точек эталонной выборки в каждой. Результат применения описанной процедуры разбиения для тестового примера приведён на Рис7.

## СПИСОК ЛИТЕРАТУРЫ

1. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. *Прикладная статистика: классификация и снижение размерности.* М.: Финансы и статистика, 1989.
2. Классификация, URL: http://www.machinelearning.ru/wiki/index.php?title=Классификация (дата обращения: 10.04.18)
3. Черезов Д. С.,Тюкачев Н. А. Обзор основных методов классификации и кластеризации данных. *Вестник ВГУ, серия: Системный анализ и информационные технологии.* 2009, №2, стр.25-29.
4. Нестеренко В. А. Классификация данных на основе функций распределения. *Современные методы и проблемы теории операторов и гармонического анализа и их приложения - VIII*: материалы докладов международной конференции, Ростов-на-Дону, 22-27 апреля 2018, стр.124-125.
5. Критерий хи-квадрат, URL: http://www.machinelearning.ru/wiki/index.php?title=Критерий_хи-квадрат (дата обращения: 10.04.18)

# The method of data classification and distribution functions

**Nesterenko V.A.**

The article proposes a new method of data classification. The method is based on a comparison of the test and training samples. Based on the results of the comparison, it is concluded that the objects of the test sample belong to the class represented by the training sample. A criterion for the possibility of an object to enter the desired class is formulated. The results of practical application of the proposed method are given. This method can be used to classify data in Data Mining tasks.

**KEYWORDS:** classification, classification of data, distribution function, Chi-square statistics.