УДК 681.3.06 + 510.63

Бутенков С.А.

НЕМЕТРИЧЕСКИЙ ПОДХОД В ЗАДАЧАХ ГРАНУЛЯЦИИ ДАННЫХ

ООО "НИЦ супер-ЭВМ и нейрокомпьютеров",
Россия, Таганрог, пер. Итальянский 106, 347900

Butenkov S.A.

NON-METRIC APPROACH FOR THE DATA GRANULATION

Scientific Research Center of Super-computer and Neuro-computer,
Russia, Taganrog, Italyanskij st., 106, 347900

Аннотация. В работе рассматривается построение математической модели для широко распространенных в различных областях данных не-метрической природы (из гуманитарных наук, химии и т.д.). Введена модель и связанные с ней определения, позволяющие строить эффективные алгоритмы гранулирования, анализа и обработки данных в гранулированной форме.

Ключевые слова: мягкие вычисления, информационная грануляция, гранулированные вычисления, анализ данных.

Abstract. In this paper we describe the use of mathematic model for the wide represented kind of non-metric data (from the pertaining to the humanities domains, chemistry etc.). The new type of model was introduced and described. This model provides the efficient data granulation, analysis and processing algorithms over the granulated data.

Key words: soft computing, information granulation, granular computing, data mining.

Вступление.

Систематизация и классификация различных видов данных лежат в основе аналитической, и, более того, интеллектуальной деятельности. Задача моделирования этих процессов является ключевой для широкого круга

практических задач в области компьютерной обработки и распознавания данных, принятия решений и т.д. Широкий класс задач, связанных с классификацией и сопряженными с ней подзадачами, выделяется в довольно эклектичную область методов, называемую *гранулированными вычислениями* (*Granular Computing, GrC*). Необходимость в создании новых методов и алгоритмов в данной области связана с тем, что объемы накопленных данных различных типов угрожающе растут с каждым годом. Теоретической основой методов GrC является использование различных классов метрик в евклидовом пространстве моделей данных. Однако существует множество типов данных, не допускающих формальное представление в виде моделей в метрических пространствах. Таковы, например, данные о цвете объектов, широко используемые в задачах обработки и анализа изображений.

В настоящей работе обобщаются теоретические результаты ряда работ, связанных с введением нового типа методов GrC, применимых к более широким классам данных, чем популярные алгоритмы Granular Computing и Data Mining, распространенные в настоящее время. Эти методы развивают основы общей Теории информационной грануляции (ТИГ), введенной в работах L.A. Zadeh.

Обзор литературы.

Методы кластеризации, в силу своей важности, развивались для широкого спектра прикладных задач [1]. Качественное обновление и значительное увеличение релевантности к важным прикладным задачам в области кластеризации и классификации было достигнуто с введением в работах L.A. Zadeh *теории нечетких множеств* (*Fuzzy Sets, FS*) [2], которая в его последующих работах была развита в общую *теорию информационной грануляции* (ТИГ) [2], [3]. На этой теоретической базе развились различные модификации моделей FS в задачах обработки данных, такие как Shadowed Sets [4], Rough Sets [5] и т.д. [6].

Концептуальной основой методов GrC является введенный L.A. Zadeh *руководящий принцип мягких вычислений*, согласно которому следует

использовать неточность, неопределенность в исходных данных для достижения наглядности обработки, нечувствительности к ошибкам, низкой стоимости решения и лучшего соответствия с реальностью внешнего мира [2].

Однако классические методы кластеризации не могут применяться для таких данных. Более новые методы GrC также ориентированы на модели исходных данных на основе евклидова пространства с метрикой. Следовательно, актуальной является задача расширения методов GrC на предметные области данных не-метрической природы [7].

Формализация модели исходных данных.

В теории информационных процессов совокупности смежных в пространстве или во времени значений данных принято считать *образами*, которые можно трактовать как упорядоченные наборы вещественных чисел. Согласно методам, теории анализа информации, такие наборы обычно описываются с помощью точек *абстрактного пространства* [9]. Введем необходимые для формализации базовых понятий определения.

Пусть K – произвольное *числовое поле*, а $n \in \mathbb{N}$ – число. Тогда n-мерным *числовым вектором* над полем K мы будем называть любой *кортеж*, составленный из n чисел поля K. Элементы числового поля, из которых составлен кортеж, называют *координатами* вектора. n-мерным *числовым пространством* над полем K называется совокупность всех n-мерных числовых векторов над этим полем [10]. Для формализации понятия абстрактного векторного пространства откажемся от ограничения природы векторов этого пространства в следующих определениях.

Определение 1. Любое множество $\mathbf{L}$ произвольных элементов называется *векторным пространством* над данным числовым полем K, если:

1. Имеется некоторая операция, ставящая в соответствие каждой паре элементов $\mathbf{a},\mathbf{b} \in \mathbf{L}$ некоторый элемент $\mathbf{c} = \mathbf{a} + \mathbf{b}$, $\mathbf{c} \in \mathbf{L}$, называемый *суммой*.

2. Имеется вторая операция, ставящая в соответствие каждому элементу $\mathbf{a} \in \mathbf{L}$ и каждому числу $k \in K$ $\mathbf{c} = k\mathbf{a}$, $\mathbf{c} \in \mathbf{L}$.

3. Обе операции удовлетворяют следующим аксиомам:

I. Для любых $\mathbf{a},\mathbf{b},\mathbf{c} \in \mathbf{L}$ имеют место соотношения:

а) $\mathbf{a}+\mathbf{b}=\mathbf{b}+\mathbf{a}$ (коммутативность),

б) $(\mathbf{a}+\mathbf{b})+\mathbf{c}=\mathbf{a}+(\mathbf{b}+\mathbf{c})$ (ассоциативность).

II. В $\mathbf{L}$ существует $\mathbf{0}$ (нулевой элемент) такой, что $\mathbf{a}+\mathbf{0}=\mathbf{a}$ для всех $\mathbf{a} \in \mathbf{L}$.

III. Для всех $\mathbf{a} \in \mathbf{L}$ существует такой элемент $-\mathbf{a}$, называемый *противоположным* для $\mathbf{a}$, что $\mathbf{a}-\mathbf{a}=\mathbf{0}$.

IV. Для любых $\mathbf{a},\mathbf{b} \in L$ и любых чисел k_1 и k поля K имеют место соотношения:

а) $k_1(k_2\mathbf{a})=(k_1k_2)\mathbf{a}$, б) $(k\ +k_2)\mathbf{a}=k_1\mathbf{a}+k_2\mathbf{a}$, в) $k_1(\mathbf{a}+\mathbf{b})=k_1\mathbf{a}+k_1\mathbf{b}$.

V. Для любого $\mathbf{a} \in \mathbf{L}$ имеет место $\mathbf{1a}=\mathbf{a}$.

В рамках приведенного определения любой элемент рассматриваемой совокупности называют *абстрактным вектором* [10].

Рассмотрим теперь m-мерное векторное пространство, на m координатных осях которого определены собственные *единичные векторы* (*орты*) $\mathbf{e}_1,\mathbf{e}_2,...,\mathbf{e}_m$ и разложение вектора в виде $V=\sum_{i=1}^{m} v_i\mathbf{e}_i$, где v_i – координаты абстрактного вектора. Если в рассматриваемом векторном пространстве не представляется возможным сравнение длин $\mathbf{e}_1,\mathbf{e}_2,...,\mathbf{e}_m$, его называют *аффинным* векторным пространством. В противном случае, т. е. если возможно найти общую метрику для $\mathbf{e}_1,\mathbf{e}_2,...,\mathbf{e}_m$, пространство называется *метрическим* векторным пространством.

Для задач анализа данных проблема построения модели данных в виде метрического пространства является ключевой при применении стандартных методов анализа многомерных данных. Во многих физических и, особенно, социологических, химических, биологических и т.п. приложениях имеет место *несоизмеримость* множеств анализируемых параметров. Меры на них вводятся искусственно и не всегда корректно [1]. Строгим является подход, основанный на использовании аффинных моделей данных [7]. Аффинное векторное пространство позволяет изучать общие свойства фигур, не изменяющиеся при

произвольном преобразовании системы координат [13]. При такой постановке мы сможем строить методы анализа произвольных данных с наибольшей общностью [12]. Для аффинного пространства введем понятие *определителя*. Отметим, что в дальнейшем номер вектора мы будем выносить влево, чтобы не путать с индексами координат вектора.

Определение 2. Определителем порядка n называется функция от n n-мерных векторов вида $F\left({}^1\mathbf{a},{}^2\mathbf{a},...,{}^n\mathbf{a}\right) = \left|{}^1\mathbf{a},{}^2\mathbf{a},...,{}^n\mathbf{a}\right|$, удовлетворяющая следующим аксиомам:

А) $F\left({}^1\mathbf{a},{}^2\mathbf{a},...,{}^n\mathbf{a}\right)$ линейна по отношению к каждому аргументу.

Б) Если среди векторов ${}^1\mathbf{a},{}^2\mathbf{a},...,{}^n\mathbf{a}$ есть хотя бы пара линейно зависимых, то определитель равен нулю.

В) $\left|\ \mathbf{e},{}^2\mathbf{e},...,{}^n\mathbf{e}\right| = 1.$

Общие алгебраические свойства определителя, вытекающие из данного определения, приведены в [10]. С геометрической точки зрения определитель в n-мерном аффинном векторном пространстве представляет собой некоторую меру на многограннике, построенном на векторах ${}^1\mathbf{a},{}^2\mathbf{a},...,{}^n\mathbf{a}$, имеющую знак (ориентированная площадь, объем и т.д.) [13]. В ряде наших работ введены модели в виде определителей специального вида, позволяющие придать исходным векторам данных геометрический смысл и строить на них меру сходства (необходимую для кластеризации или грануляции) без использования метрики [14], [15].

Грануляция многомерных данных.

Введем теперь необходимые определения теории грануляции и предлагаемую в наших работах модель гранулированного представления объектов в аффинных векторных пространствах [8].

В теории информационной грануляции (ТИГ) L.A. Zadeh *информационной гранулой* называется подмножество универсума, на котором определено отношение сходства, неразличимости и т.п. [1]. Множество гранул, которое

содержит все объекты универсума, называется *гранулированием* универсума. Введем ряд формализованных определений, уточняющих мереологию подхода L.A. Zadeh.

Определение 3. Разбиение конечного универсума **G** – это конечное множество подмножеств ${}^{i}G \in \mathbf{G}$, $i = 1,...,n$ (*атомарных гранул*), удовлетворяющих следующим аксиомам:

1. ${}^{i}G \neq \varnothing, \quad i = 1,...,n;$

2. ${}^{i}G \cap {}^{j}G = \varnothing$ при $i \neq j; \quad i = 1,...,n; \; j = 1,...,n;;$

3. $\bigcup_{i} {}^{i}G = \mathbf{G}, \quad i = 1,...,n.$

Каждое подмножество разбиения называется *гранулой эквивалентности*. Подмножество ${}^{i}G \subseteq \mathbf{G}$ называется *составной гранулой* (не элементарной) если оно представляет собой объединение атомарных гранул определения 4 [8]. Переход от исходного векторного пространства к гранулированному представлению (покрытию или разбиению) универсума выполняется на основе ряда базовых понятий.

Определение 4. Покрытие τ конечного универсума U – это конечное множество подмножеств ${}^{i}G$, удовлетворяющих аксиомам 1 и 3.

Определение 5. Разбиение π (или покрытие τ) называется *конъюнктивным* разбиением (покрытием) если каждый класс эквивалентности из π (τ) – составная гранула.

Определение 6. Разбиение π_1 есть *уточнение разбиения* π_2 (или π_2 есть обобщение разбиения π_1), обозначаемое как $\pi_1 \prec \pi_2$, если каждая гранула из π_1 содержится в некоторой грануле из π_2. Покрытие τ_1 есть *уточнение покрытия* τ_2 (или τ_2 есть обобщение покрытия τ_1), обозначаемое как $\tau_1 \preceq \tau_2$, если каждая гранула из τ_1 содержится в некоторой грануле из τ_2.

Таким образом, основу гранулированных вычислений (GrC) составляют методы построения изоморфных систем гранул на основании исходных данных с применением уточнений на каждом этапе моделирования [4].

В работах L.A. Zadeh введено общее определение *декартовой* информационной гранулы, которое можно распространить на общий случай аффинных пространств.

Определение 7. Пусть заданы произвольные гранулы ${}^1G,...,{}^nG$ размерности 1 для переменных $U_1,...,U_n$ соответственно, тогда их декартово произведение $G_n = {}^1G \times \cdots \times {}^nG$ – это *декартова гранула* размерности n.

Кластер A точек аффинного пространства, имеющий проекции на оси соответственно $proj_{x_1}A$ и $proj_{x_2}A$ покрывается их декартовым произведением – информационной гранулой G_2^+ размерности $n=2$. в ТИГ L.A. Zadeh этот процес называется *инкапсуляцией информации*. Введем следующее определение, лежащее в идеологии работы [3]:

Определение 8. Декартова гранула G^+ определяемая как $G^+ = G_x \times G_y$, *инкапсулирует* исходную произвольную гранулу G в том смысле, что является точной верхней гранью декартовых гранул, которые содержат G.

С геометрической точки зрения декартова гранула строится как декартово произведение подмножеств на осях векторного пространства. Для случая $n=2$ определения, связанные с покрытием множеств декартовыми гранулами [8] иллюстрирует рис.1.

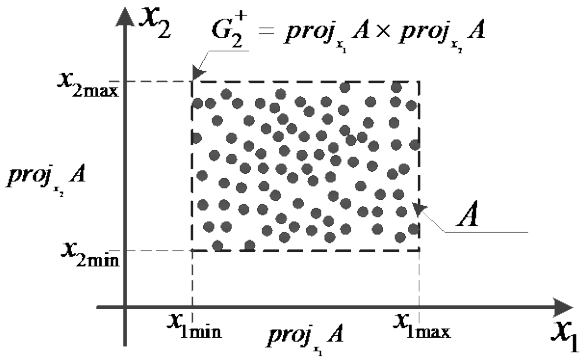

Рис. 1. Покрытие произвольного подмножества плоскости A декартовой гранулой G_2^+ (инкапсуляция данных).

Таким образом, G_n^+ может использоваться как верхняя аппроксимация кластера A в аффинном пространстве размерности n [14]. Модификация

гранулы с помощью операций уточнения может быть инструментом оптимизации исходного представления гранулы G_n^+ в конкретных задачах грануляции [8]. Введем тепрерь математический аппарат для реализации основных определений ТИГ в аффинном пространстве.

Модель декартовой гранулы в аффинном векторном пространстве.

Новым подходом в представлении многомерных данных является использование грассманновых элементов различной размерности [8], реализующих принцип пространственной грануляции данных для сложных объектов [4]. В декартовых координатах гранула в пространстве размерности n представляется в виде матрицы из векторов аффинного пространства ${}^i\mathrm{X}$, $i=1,...,n$, называемой *грассманновым элементом* размерности n [13]:

$$G_n=\begin{pmatrix} {}^1x_1 & {}^2x_1 & \cdots & {}^nx_n & 1 \\ {}^1x_2 & {}^2x_2 & \cdots & {}^nx_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & 1 \\ {}^1x_n & {}^2x_n & \cdots & {}^nx_n & 1 \\ {}^1x_{n+1} & {}^2x_{n+1} & \cdots & {}^nx_{n+1} & 1 \end{pmatrix}. \qquad (1)$$

Вычисление параметров модели (1) выполняется с помощью покрытия по определению 4 с отысканием покрывающей гранулы как верхней грани подмножества по рис.1. Для этого примера (при $n=2$) можно записать метод построения модели (1) для инкапсулирующей декартовой гранулы G_2^+ по m точкам$\left({}^ix_1,{}^ix_2\right)$, $i=1,2,...,m$ исходной (не декартовой) гранулы $proj_{x_1}A$ и $proj_{x_2}A$ в виде:

$$G_2^+(A)=\begin{pmatrix} \min\left({}^ix_1\right) & \min\left({}^ix_2\right) & 1 \\ \min\left({}^ix_1\right) & \max\left({}^ix_2\right) & 1 \\ \max\left({}^ix_1\right) & \max\left({}^ix_2\right) & 1 \end{pmatrix},\ i=1,2,...,m. \qquad (2)$$

Аналогичные (2) формулы можно получить для произвольной размерности пространства n [8]. Возможны уточнения покрытия по определениям 5, 6, улучшающие качество инкапсуляции данных.

Для реализации инкапсуляции двух декартовых гранул (1) одной размерности (например, $n=2$) используется аналогичная (2) формула:

$$G_2^+\left({}^iG_2, {}^jG_2\right)=\begin{pmatrix} \min\left({}^ix_1, {}^jx_1\right) & \min\left({}^ix_1, {}^jx_1\right) & 1 \\ \min\left({}^ix_2, {}^jx_2\right) & \max\left({}^ix_2, {}^jx_2\right) & 1 \\ \max\left({}^ix_3, {}^jx_3\right) & \max\left({}^ix_3, {}^jx_3\right) & 1 \end{pmatrix}. \quad (3)$$

Согласно определению 2, мы можем вычислить определитель такой модели в аффинном векторном пространстве. Его значение определяет некоторую *меру* на грануле, позволяющую создать методы грануляции и манипулирования гранулированными представлениями данных [14].

Меры множеств и информационная грануляция.

Мера является математической основой обработки и особенно анализа данных. Дадим общее определение, охватывающее широкий класс мер на гранулированных пространствах.

Определение 9. Неотрицательная функция множества $m:\mathbf{P}(X)\to\mathbb{R}$, называется *мерой* множества, если удовлетворяет следующим аксиомам:

1. $A\subseteq X \Leftrightarrow m(A)\geq 0$;

2. $m(\varnothing)=0$;

3. если $A,B\in\mathbf{P}(X)$, то $m\left(A\cup B\right)=m(A)+m(B)-m(A\cap B)$.

Здесь $\mathbf{P}(X)$ – множество всех подмножеств X, $\mathbb{R}$ – множество действительных чисел. Рассмотрим теперь ряд мер, связанных со специфическими множествами в пространстве $\mathbf{G}$ гранулированного представления данных [8]. Согласно [4] для единичной гранулы $G\in\mathbf{G}$ задается *мера общности* вида $Glob(\mathrm{G})=m(\mathrm{G})/m(\mathbf{G})$, которая определяет относительный размер гранулы G. Для двух гранул ${}^iG,{}^jG\in\mathbf{G}$ задается *мера соответствия* в виде $AS({}^i\mathrm{G},{}^j\mathrm{G})=m({}^i\mathrm{G},\cap{}^jG)/m(\mathbf{G})$. *Мера покрытия* для ${}^iG,{}^jG\in\mathbf{G}$ задается в виде $\mathrm{CV}({}^i\mathrm{G},{}^j\mathrm{G})=m({}^i\mathrm{G},\cap{}^jG)/m({}^j\mathrm{G})$.

Метод реализации мер $m(\mathrm{G})$, удовлетворяющих определению 3, задается для конкретных предметных областей применения гранулирования [12].

Введенная здесь модель декартовых гранул допускает вычисление целого спектра мер на гранулах G_n на основе миноров определителя модели (1) в аффинном пространстве. Так, три базовые меры на двумерной грануле ${}^2G$, имеющие очевидный геометрический смысл в аффинном пространстве [8], задаются уравнениями определителей на матрицах моделей:

$$\eta^1(G_2)=\begin{vmatrix} {}^1x_1 & 1 \\ {}^1x_2 & 1 \end{vmatrix},\ \eta^2(G_2)=\begin{vmatrix} {}^2x_1 & 1 \\ {}^2x_2 & 1 \end{vmatrix},\ \eta^3(G_2)=\begin{vmatrix} {}^1x_1 & {}^2x_2 & 1 \\ {}^1x_2 & {}^2x_2 & 1 \\ {}^1x_3 & {}^2x_3 & 1 \end{vmatrix}. \quad (4)$$

Аналогичным образом можно получить выражения для мер на декартовых гранулах G_n для произвольной размерности n [9]. В [14] показано, что функции гранул $\eta^1(G_n),\eta^2(G_n),\eta^3(G_n)\in\mathbb{R}$ (5) удовлетворяет аксиомам определения 10. На основе определения 10 и меры из (4) общее выражение меры сходства двух грассманновых элементов ${}^iG_n,{}^jG_n\in\mathbf{G}$, инкапсулируемых $G_n^+\left({}^iG_n,{}^jG_n\right)\in\mathbf{G}$ [14], представляется в виде:

$$\mathrm{SIM}({}^i\mathrm{G}_n,{}^j\mathrm{G}_n)=\eta^3\left(G_n^+\right)\Big/\left(\eta^3\left({}^iG_n\right)+{}^3\left({}^iG_n\right)\right). \quad (5)$$

В [14] показано, что функция гранул $\mathrm{SIM}({}^i\mathrm{G}_n,{}^j\mathrm{G}_n)\in\mathbb{R}$ удовлетворяет аксиомам определения 10 и является мерой в аффинном пространстве данных размерности n. Отметим, что применение (5) в задачах гранулирования данных не требует существования нормы применяемого пространства в отличие от классических методов кластеризации [1] и грануляции [5].

Результаты. Обсуждение и анализ.

Полученные в работе теоретические результаты развивают основные положения Теории информационной грануляции применительно к анализу данных не-метрической природы. Для их применения не требуется введения нормы и евклидова пространства данных, что не всегда возможно для произвольных типов данных. Предложенная модель гранулы в аффинном пространстве допускает преобразование координат, при этом нет необходимости вводить новые нормы. Такие свойства позволяют

распространить методологию грассманновых элементов на произвольные системы криволинейных координат, что расширяет возможности моделирования типов данных (например, цветовых моделей). В наших работах показано, что использование моделей (1)–(5) позволяет строить эффективные жадные алгоритмы обработки и анализа данных, а также нейросетевые структуры, решающие эти задачи. Эти проблемы будут рассмотрены в следующих работах.

Заключение и выводы.

Были рассмотрены различные подходы к построению математических моделей данных реального мира. Введена базовая методология представления данных в аффинном (не-метрическом) пространстве, на основании которой развиты методы теории информационной грануляции и гранулированных вычислений для анализа данных. Этим методам придано концептуальное единство.

Были получены модели данных, допускающие эффективную реализацию для задач анализа и обработки данных не-метрической природы. Для введенных математических моделей предложены меры, позволяющие анализировать и обрабатывать такие типы данных без использования метрических характеристик.

Литература:

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. - 336 с.: ил.

2. Zadeh L.A. Fuzzy Sets and Information Granularity // Advances in Fuzzy Set Theory and Applications, M. Gupta, R. Ragade, and R. Yager, Eds. Amsterdam, The Netherlands: North-Holland, 1979, pp. 3–18.

3. Zadeh L.A. Toward a Theory of Fuzzy Information Granulation and its centrality in human reasoning and fuzzy logic // Fuzzy Sets Syst., vol. 90, pp. 111–127, 1997.

4. Pedrysz W. Granular Computing – the emerging paradigm // Journal of Uncertain Systems, Vol.1, No.1, pp.38-61, 2007

5. Pawlak Z. Granularity of Knowledge, Indiscernibility and Rough Sets // Proceedings of 1998 IEEE International Conference on Fuzzy Systems, 106-110, 1998.

6. Yao Y. Granular Computing: Basic Issues and Possible Solutions // Proc. of the 5th Joint Conference on Information Sciences, 2000. Pp. 186-189.

7. Бутенков С.А., Жуков А.Л. Информационная грануляция на основе изоморфизма алгебраических систем // Сб. трудов Международной алгебраической конференции, посвященной 80-летию со дня рождения А.И. Кострикина, Нальчик, 12-18 июля 2009 г., с. 206-209.

8. Бутенков С.А., Жуков А.Л. Гранулирование геометрических данных в задачах автоматизированного проектирования // Журнал “Известия ЮФУ. Технические науки”, №9, 2008, с. 87-92.

9. Zilles S.N. Introduction to Data Algebras, Lect. Notes Comp. Sci. 86, 1980, 248-272.

10. Общая алгебра. Т. 1, 2. / В.А. Артамонов, В.Н. Салий, Л.А. Скорняков и др. Под общ. ред. Л.А. Скорнякова. - М.: Наука, 1991.

11. Колмогоров А.Н., Фомин С.В. Элементы теории функций и функционального анализа. - М.: Наука, 1989. - 544 с.

12. Бутенков С.А., Нагоров А.Л., Бесланеев З.О. Геометрический подход к построению моделей данных на основе теории грануляции // Вестник Дагестанского гос. тех. университета.– Махачкала: Изд.-во ДГУ, 2014, №1, т. 32, с. 47-55.

13. Klein F. Elementarmathematik vom Hoheren Standpunkte Aus Erster Band, Berlin, Verlag von Julius Springer, 1924.

Статья отправлена: 12.12.2015 г.