

## *ЭНТРОПИЯ В ДЕРЕВЕ ПРИНЯТИЯ РЕШЕНИЙ, ПОЛУЧЕНИЕ КОТОРОГО ПРИБЛИЖЕНО К ЕСТЕСТВЕННОМУ ИНТЕЛЛЕКТУ*

**О.Б. ПОПОВА**

*Кубанский государственный технологический университет,  
350072, Российская Федерация, Краснодар, ул. Московская, д. 2,  
электронная почта: popova\_ob@mail.ru*

Актуальной задачей является классификация знаний заданной предметной области, где они представляются не закодированной определённым образом информацией, а способом близким к естественному интеллекту, который структурирует получаемые знания по иному принципу. Классифицировать известные ответы на вопросы необходимо так, чтобы решить текущую задачу. Так к пониманию знания подходит новый метод формирования дерева принятия решений, приближенный к естественному интеллекту. В статье рассматривается, как энтропия связана с появлением знаний, классификацией предыдущих знаний и с определениями, используемыми в дереве принятия решений. Последнее необходимо для сравнения алгоритма построения дерева принятия решений, приближенного к естественному интеллекту, с традиционными методами. Получена зависимость энтропии от свойств подмножества элементов множества. Установлено, что для уменьшения энтропии и классификации необходимо определять свойства для подмножеств, величина которых постепенно уменьшается. Условием прекращения деления подмножества на части является число элементов подмножества меньше 2. Рекурсивное повторение простейших действий делает алгоритм эффективным и простым для реализации бинарным деревом системы вопросов и ответов, создаёт простые продукционные правила.

**Ключевые слова:** энтропия, естественный интеллект, знания, продукционное правило, дерево поддержки принятия решений, бинарное дерево системы вопросов и ответов.

Сегодня актуальной задачей является классификация знаний для заданной предметной области. Знания представляются уже в другой форме, отличной от общепринятой в теории информации и в интеллектуальных информационных системах, где единичное знание представляется как закодированная определённым образом информация. Такое представление позволяет реализовать решение с помощью вычислительной техники. Хотя человек привык представлять знание так – «это обладание проверенной информацией (ответами на вопросы), позволяющей решать поставленную задачу» [1]. Поэтому естественный интеллект привык структурировать получаемые знания именно по принципу – классифицировать существующие и известные ему ответы на вопросы таким образом, что бы решить любую поступающую к нему задачу. Именно с такой позиции к пониманию знания

подходил автор статьи при получении нового метода формирования дерева принятия решений, приближенного к естественному интеллекту.

Определим как понятие «энтропия» связана с появлением новых знаний, классификацией предыдущих знаний и свяжем её с основными определениями, используемыми в построении деревьев принятия решений. Последнее необходимо для сравнения нового метода построения дерева принятия решений, приближенного к естественному интеллекту [2], с традиционными методами [3, 4].

Энтропия необходима для того, чтобы отразить способ структурирования знаний, используемый естественным интеллектом, в получении узлов дерева, а так же сформулировать алгоритм, который сравним с известными алгоритмами [3].

Известно [4], что энтропия «зависит от пропорции, в которой разделяется множество», а по «мере возрастания этой пропорции от 0 до  $\frac{1}{2}$  - симметрично убывает» (см. рис. 1). В «случае деления множества на несколько частей энтропия максимальна тогда, когда части равновеликие, и равна нулю, когда одна из частей занимает всё множество» [4, с. 35]. Для получения эффективного разделения множества на подмножества необходимо подбирать такой атрибут, который даёт наибольший прирост информации [4]. Тогда в стандартном понимании энтропии, принятом в теории информации, «листьями дерева принятия решений являются классы», чтобы «классифицировать объект ... нужно последовательно спускаться по дереву», тогда путь «от корня дерева до листьев можно трактовать как объяснение того, почему тот или иной объект отнесён к какому-либо классу» [3]. Тогда «энтропия представляет собой среднее количество битов, которые требуются, чтобы закодировать атрибут  $S$  у элемента множества  $A$ » [4, с. 35]. Тогда, если «вероятность появления  $S$  равна  $\frac{1}{2}$ , то энтропия равна 1, и нужен полноценный бит; а если  $S$  появляется не равновероятно, то можно закодировать последовательность элементов  $A$  более эффективно» [4, с.35]. Данное представление энтропии позволяет использовать теорию кодирования для получения оптимального кода бинарного дерева

принятия решений, чтобы использовать его в интеллектуальных информационных системах, например, в системах поддержки принятия решений, для решения различных задач. Такие деревья принятия решений позволяют применять к ним известные способы представления данных, которые приняты в современных базах данных и базах знаний. Это представление энтропии, информации и знаний характерен для того процесса, который хорошо можно автоматизировать с помощью вычислительной машины, но он отличается от способа структурирования информации естественным интеллектом.

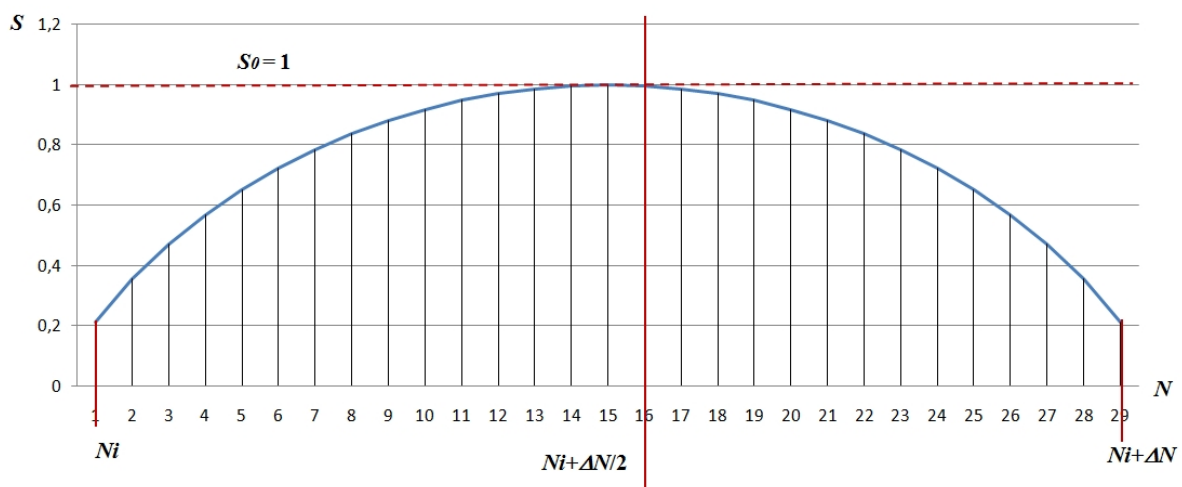


Рисунок 1 – Энтропия Шеннона, которая зависит от выбора свойства разделяющего задачи определённых типов на две части

Естественному интеллекту характерно оперировать не битами, а теми объёмами знаний, которые он усвоил, то есть понял и, следовательно, смог соответствующим образом структурировать. Тогда, в соответствии с указанными выше положениями из теории информации и определениями энтропии можем считать, что прирост знаний  $Z$  на определённую величину  $\Delta Z$  позволит уменьшить энтропию  $S$  на некоторую величину  $\Delta S$  (см. рис. 2). Это возможно, если знание  $\Delta Z$  отличается от знания  $Z$  каким-то свойством  $C$ . Свойство  $C$  можно считать атрибутом.

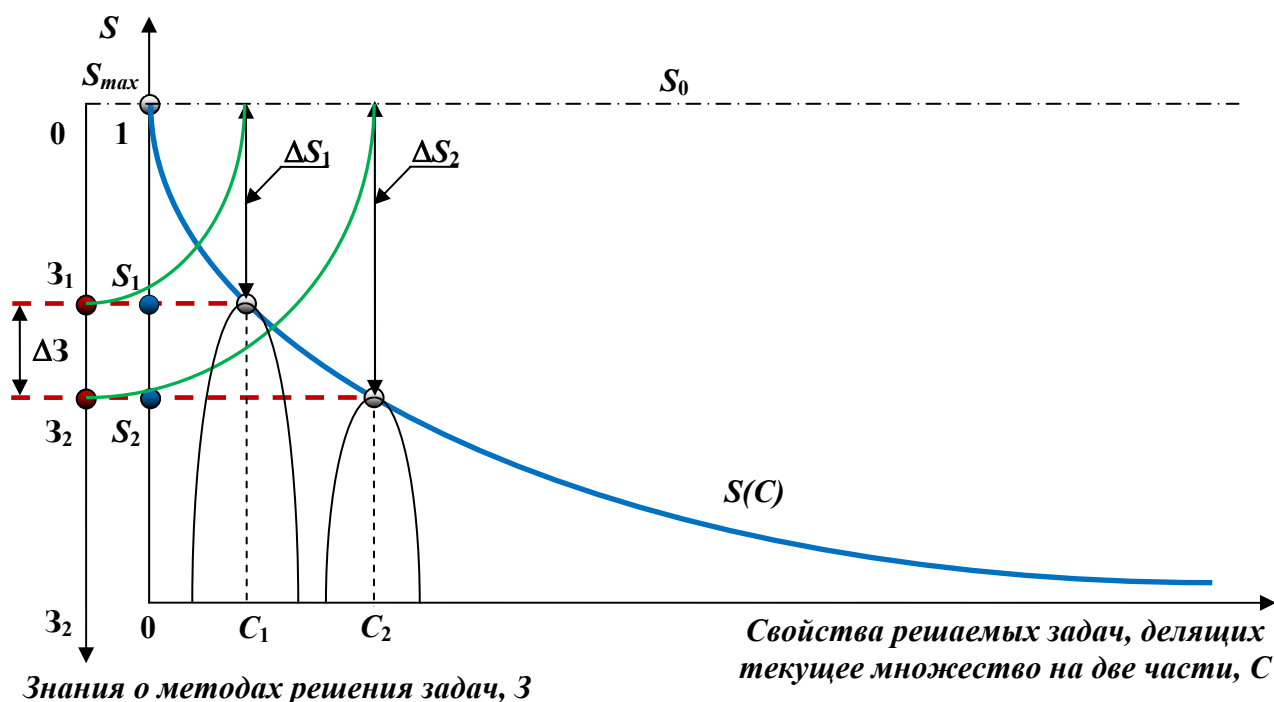


Рисунок 2 – Характер изменения максимальной энтропии от времени, которая зависит от появления нового знания, то есть новых методов решения задач нового типа

Теперь соединим известные принципы структурирования информации естественным интеллектом для ее запоминания [5] с определениями энтропии и положениями, принятыми в теории информации.

Исходя из первого принципа структурирования «информация должна быть поделена на группы и подгруппы в соответствии с определенным значимым для нас критерием» [5]. Согласно второму принципу «выделенные группы должны быть логично связаны, выстроены в необходимом порядке (по важности, по времени, по интенсивности и т.п.)» [5].

Это не противоречит утверждению из теории информации, что энтропия «зависит от пропорции, в которой разделяется множество». Поэтому она будет уменьшаться с течением времени, так как это необходимый переход от хаоса к порядку при получении удобной для запоминания логической структуры. Этот процесс будет происходить постоянно из-за появления всё новых знаний  $\Delta z$ . «Процесс деления множества данных на части, приводящий к уменьшению энтропии, можно рассматривать как производство информации» [3], где информация – это структурированные особым образом знания.

Будем считать, что до появления нового метода получения дерева принятия решений, знания из некоторой предметной области не были соответствующим образом классифицированы, то есть энтропия была максимальная. Тогда всё множество необходимо будет соответствующим образом разделять рекурсивно на две части до тех пор, пока в листьях дерева не останется один элемент. Такой способ деления множества на части позволяет учесть основные положения, полученные для энтропии в теории информации, которые приведены выше в статье, и получить оптимальную, логичную и удобную для реализации и запоминания структуру дерева принятия решений. Проще было бы разделить множество на части равные числу элементов множества, тогда за один раз деления энтропия уменьшилась до нуля. Такое возможно, если каждый элемент имел бы своё свойство. Но, как правило, для любого множества знаний некоторой предметной области существует ряд свойств, каждому из которых соответствует некоторый набор элементов из множества. В результате каждый элемент множества обладает своим набором свойств. Поэтому нужно подбирать для деления в первую очередь те свойства, которым обладает большее число элементов из множества (см. рис. 3).

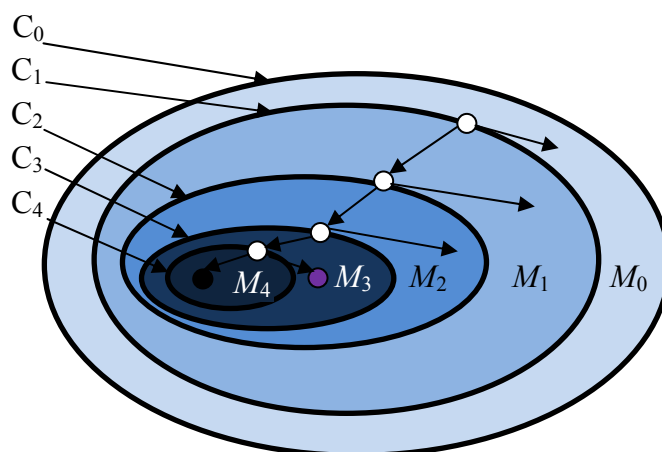


Рисунок 3 - Множество знаний некоторой предметной области со своим набором свойств

На рисунке 3 видно, что все элементы множества  $M_0$  обладают общим свойством  $C_0$ . Тогда подмножество  $M_1$  имеет свойство  $C_1$ , которое объединяет большее число элементов из исходного множества  $M_0$ . Следовательно это свойство первым разобьет исходное множество  $M_0$  на две части, уменьшив

максимальную энтропию  $S_{max}$  на величину  $\Delta S_1$ . Таким же образом будет уменьшена энтропия  $S_1$  на величину  $\Delta S_2$  делением подмножества  $M_1$  свойством  $C_2$  на две части. Этот процесс будет продолжаться пока подмножество  $M_3$ , состоящее из двух элементов (чёрный и фиолетовый шарики) не будет разделено на две части свойством  $C_4$ . Так как подмножество  $M_4$  состоит из одного элемента (чёрный шарик), то энтропия близка к нулю. Будем считать её близкой к нулю, так как есть вероятность появления с течением времени ещё нескольких элементов подмножества  $M_4$  с таким же свойствами  $C_0, C_1, C_2, C_3$  и  $C_4$ .

Рекурсивным образом повторяет те же действия с оставшимися элементами из подмножеств –  $M_0 - M_1; M_1 - M_2; M_2 - M_3$ . Так как в  $M_3$  остался всего лишь один элемент (сиреневый шарик), то разбиение на части не производится.

Теперь структурированные особым образом знания о предметной области дают наименьшее значение энтропии, но не ноль, так как знания будут продолжать генерироваться учёными. Это будет продолжаться до тех пор, пока будут появляться всё новые и новые научные и прикладные задачи, требующие решения. Задачи могут быть разной степени сложности и затрагивать темы по уже известным и классифицированным знаниям. Поэтому к разбиению множества на части добавляется ещё два действия – переход по дереву принятия решений с поиском места потенциального добавления элементов в нужное подмножество и дальнейшего его разбиения на части указанным выше способом; добавления нового корня дерева, если появились элементы со свойствами или свойством отличными от уже имеющихся свойств. Последнее действие похоже на деление начального множества на две части свойством, которое объединяет наибольшее число элементов подмножества. Если появилось несколько элементов со свойством  $C_{-1}$ , то наибольшим подмножеством (см. рис. 3) будет  $M_0$  со свойством  $C_0$ , которое разобьёт всё множество на две части. Первой частью будет уже классифицированные элементы, которые войдут в дерево без изменений. А второй частью –

подмножество из нескольких элементов со свойством  $C_{-1}$ , которые будут рекурсивно поделены аналогичным образом. Следовательно, при повторении такой ситуации следующим свойством, которое создаёт новый корень дерева, будет  $C_{-1}$ . При получении нового корня дерева энтропия постепенно уменьшается и стремится к нулю.

Определим характер зависимости  $S(C)$ . На рисунке 4 видно, что при первичном делении множества  $M_0$  на две части свойством  $C_2$  график имеет форму прямой для элементов подмножества  $M_1$  с одинаковым свойством  $C_2$  – это отрезок с точками 1-2. При делении этого подмножества  $M_1$  на две части свойством  $C_3$  получится ломаная линия с точками 1-3-2. Таким образом, при дальнейшем делении подмножеств на две части ломаная линия приобретает прогиб и становится более сглаженной. Для изображённого на рисунке 4 примера – это ломаная линия с точками 1-9-8-3-7-4-5-6-2. С течением времени новые знания сделают данную ломаную линию очень сглаженной по форме близкой к логарифмической функции. Следовательно, зависимость  $S(C)$  можно сразу же изображать в виде такого графика (см. синяя логарифмическая кривая)

$$-\sum_{i=1}^2 (p_i * \log_2 p_i) = -\frac{N_i}{N} * \log_2 \frac{N_i}{N} - \frac{N - N_i}{N} * \log_2 \frac{N - N_i}{N}, \quad (1)$$

где  $N_i$  – это число элементов множества со свойством  $C_i$ ;

$N$  – максимально возможное число элементов множества.

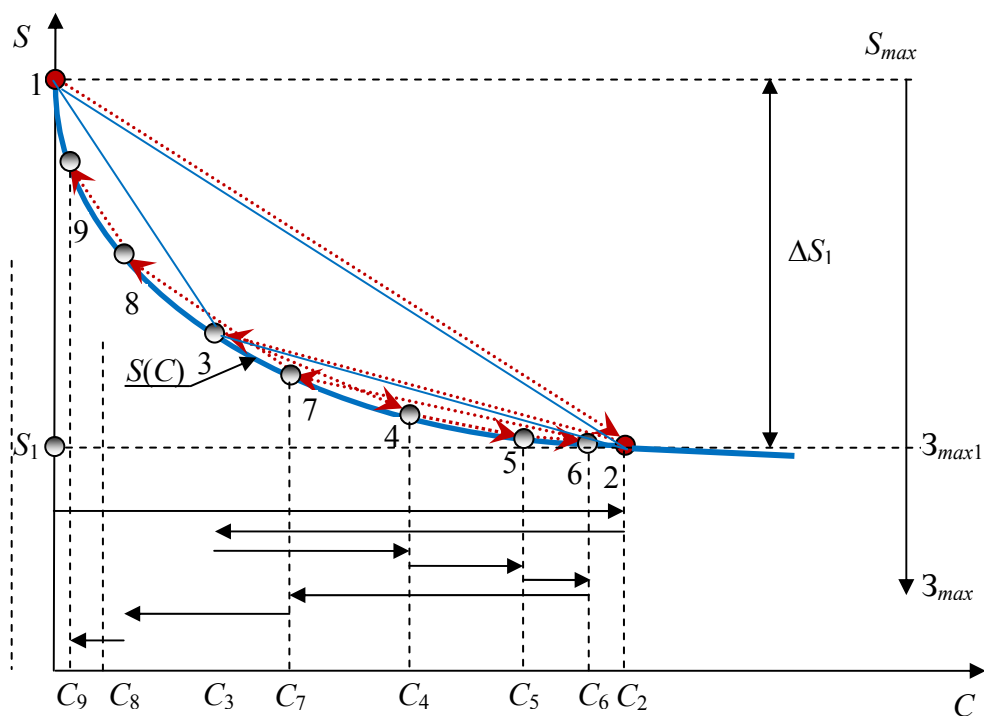


Рисунок 4 – Характер изменения зависимости  $S(C)$ , в процессе первоначального разбиения множества рекурсивно на части свойствами  $C_1, C_2, C_3 \dots, C_9$

Таким образом, график зависимости  $S(C)$  отражает не только текущие точки деления исходного множества на части, но и потенциально возможные. Формула (1) соответствует энтропии Шеннона. Если для стандартного применения энтропии в общепринятом алгоритме получения дерева принятия решений эта формула играет определённую роль при вычислении значений для предикатов заданных атрибутов для получения промежуточных узлов и листьев дерева, то для нового метода, который приближен к естественному интеллекту, расчёт энтропии не нужен.

Полученная зависимость  $S(C)$  позволяет увидеть следующие связи:

- 1) между выбором текущего корня дерева и свойством, которое объединяет наибольшее число элементов подмножества;
- 2) между выбором текущего промежуточного узла дерева и свойством, которое объединяет наибольшее число элементов текущего промежуточного подмножества.

Запишем алгоритм построения дерева принятия решений, с учётом полученных выше выводов для не классифицированного множества:



*{Подпрограмма создающая корень дерева принятия решений}*

**Создание корня дерева принятия решений для множества  $(M_0, C, M_i)$ ;**

$$M_{i-1} = M_0;$$

$C$  = определяем свойство, которым обладает подмножество  $M_i < M_0$ ;

*{Это корень дерева, на который нужно сохранить указатель}*

Создаём элемент со свойством  $C$ ;

*{Подпрограмма разбивающее множество на две части}*

**Создание текущего корня дерева принятия решений для множества  $(M_i)$ ;**

$$M_{i-1} = M_i;$$

$N$  = определяем число элементов подмножества  $M_{i-1}$ ;

Если  $N \geq 2$ , то нужно выполнить:

$M_i$  = следующее по убыванию числа элементов подмножество;

$C$  = определяем свойство, которым обладает множество  $M_i$ ;

Создаём промежуточный элемент дерева принятия решений со свойством  $C$ ;

Соединяем с деревом принятия решений *{используя указатели}*;

$$M = M_{i-1} - M_i;$$

*{Переход вниз влево, используя указатели}*

**Создание текущего корня дерева принятия решений для множества  $(M_i)$ ;**

*{Переход вниз вправо, используя указатели}*;

**Создание текущего корня дерева принятия решений для множества  $(M)$ ;**

Если  $N = 1$ , то нужно выполнить для этого элемента множества:

Создаём лист дерева принятия решений;

Соединяем с деревом принятия решений *{используя указатели}*;

*{Основная программа}*

$C_0$  = определяем свойство, которое характерно для всех элементов  $n$  множества  $M_0$ ;

**Создание корня дерева принятия решений для множества  $(M_0, C, M)$ ;**

*{Повторяем рекурсивно с разными значениями  $M_i$  и  $M_{i-1} - M_i$ , где  $i = 1, 2, 3, \dots, n$ }*

**Создание текущего корня дерева принятия решений для множества  $(M)$ ;**

Тогда для добавления новых элементов в множество  $M_0$ , необходимо использовать две указанные в алгоритме подпрограммы, которые помогут найти место для добавления элементов в дереве принятия решений или создать новый корень дерева, сохранив оставшуюся структуру не тронутой.

Можно сделать следующий вывод:

1. Необходимо определять свойства для подмножеств, величина которых постепенно уменьшается. Свойства не должны повторяться.

2. Условием прекращения деления подмножества на две части является число элементов в множестве меньше 2.

3. Возможность рекурсивного повторения простейших действий делает алгоритм удобным для реализации современными методами программирования, используя бинарное дерево [6].

4. Такое разбиение позволяет формулировать простые продукционные правила [7]:

***Если***

*свойство = параметр*

***то***

*переход к соответствующему антецеденту*

***в противном случае***

*переход к другому соответствующему антецеденту.*

5. С таким способом разбиения можно эффективно использовать разработанное автором бинарное дерево системы вопросов и ответов [2, 6].

6. Правила получения элементов бинарного дерева системы вопросов и ответов, которые разработал автор [8, 9], совпадают с алгоритмом построения дерева принятия решений, с учётом полученных в этой статье теоретических выводов.

Теория была проверена на практическом примере, где рассматривалась предметная область – методы оптимизации, которые стали элементами множества  $M_0$ . Для этого был проведён системный анализ процесса оптимизации [10-15] и получен новый способ получения дерева принятия решений.

Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда № 16-03-00382 от 18.02.2016 в рамках темы

“Мониторинг исследовательской деятельности образовательных учреждений в условиях информационного общества”.

#### ЛИТЕРАТУРА

1. Знание – [Электронный ресурс] - Режим доступа: <https://ru.wikipedia.org/wiki/Знание>.
2. Попова О.Б. Новые метод усиления интеллекта и способ представления дерева принятия решений, которые приближены к естественному интеллекту // Научные труды Кубанского государственного технологического университета. – 2016. – №1, с. 38-47; URL: <http://ntk.kubstu.ru/file/779>.
3. Энтропия и деревья принятия решений. Data Mining, Программирование. 2016. <https://special.habrahabr.ru/kyocera/p/171759/>
4. Николенко С.И., Тулупьев А.Л. Самообучающиеся системы. – М.: МЦНМО, 2009. – 288 с.
5. Буянов Е. Урок 3. Структурирование информации. [Электронный ресурс] - Режим доступа: <https://4brain.ru/memory/strukturirovanie.php>
6. Popova O., Popov B., Karandey V., Evseeva M. Intelligence amplification via language of choice description as a mathematical object (binary tree of question-answer system) // Procedia – Social and Behavioral Sciences. – 2015. – V. 214. – С. 897–905;
7. Попова О.Б. Представление метазнаний в интеллектуальной информационной системе, ведущей поисковые исследования // Научные труды Кубанского государственного технологического университета. – 2016. – №10, с. 224-236
8. Попова О.Б., Попов Б.К., Ключко В.И. Правила получения элементов бинарного дерева системы вопросов и ответов // Фундаментальные исследования. – 2013. – №6-1. – С. 55-59
9. Попова О.Б., Попов Б.К., Ключко В.И. Получение корня бинарного дерева системы вопросов и ответов // Современные проблемы науки и образования. – 2013. – №3. – С. 8
10. Попова О.Б., Попов Б.К., Ключко В.И. Проблема сокращения времени выбора методов управления большими системами (БС) // Современные проблемы науки и образования. – 2013. – №1. – С. 163

11. Попова О.Б., Попов Б.К., Ключко В.И. Анализ связей в реальной и технической системах процесса оптимизации // Международный журнал экспериментального образования. – 2013. – №10-2. – С. 405-408

12. Попова О.Б., Попов Б.К. Замена реальной системы (процесс выбора метода оптимизации) на техническую систему (программа-советчик «Оптимэль») // Современные проблемы науки и образования. – 2012. – №5. – С. 132

13. Системный анализ процесса выбора метода оптимизации информационной системы: монография / О.Б. Попова, Б.К. Попов, В.И. Ключко; ФБГОУ ВПО «Кубан. гос. технол. ун-т». – Краснодар: Издательский Дом – Юг, 2012 – 135 с.

14. Попова О.Б., Попов Б.К. Интеллектуальная информационная система выбора «Оптимэль». Патент на изобретение RUS № 2564641 от 27.05.2014.

15. Попова О.Б. Участие процесса оптимизации в развитии сложных технических систем. Деп. в ВИНТИ №257-В2010 07.05.2010.

#### REFERENCES

1. Znanie – [Elektronnyy resurs] - Rezhim dostupa: <https://ru.wikipedia.org/wiki/Znanie>.

2. Popova O.B. Novye metod usileniya intellekta i sposob predstavleniya dereva prinyatiya resheniy, kotorye priblizheny k estestvennomu intellektu // Nauchnye trudy Kubanskogo gosudarstvennogo tekhnologicheskogo universiteta. – 2016. – №1, s. 38-47; URL: <http://ntk.kubstu.ru/file/779>.

3. Entropiya i derevya prinyatiya resheniy. Data Mining, Programmirovaniye. 2016. <https://special.habrahabr.ru/kyocera/p/171759/>

4. Nikolenko S.I., Tulupev A.L. Samoobuchayushchiesya sistemy. M.: MTsNMO, 2009. 288 s.

5. Buyanov E. Urok 3. Strukturirovaniye informatsii. [Elektronnyy resurs] - Rezhim dostupa: <https://4brain.ru/memory/strukturirovaniye.php>

6. Popova O., Popov B., Karandey V., Evseeva M. Intelligence amplification via language of choice description as a mathematical object (binary tree of question-answer system) // Procedia – Social and Behavioral Sciences. – 2015. – V. 214. – S. 897–905;

7. Popova O.B. Predstavlenie metaznaniy v intellektualnoy informatsionnoy sisteme, vedushchey poiskovoye issledovaniya // Nauchnye trudy Kubanskogo gosudarstvennogo tekhnologicheskogo universiteta. – 2016. – №10, s. 224-236

8. Popova O.B., Popov B.K., Klyuchko V.I. Pravila polucheniya elementov binarnogo dereva sistemy voprosov i otvetov // Fundamentalnye issledovaniya. – 2013. – №6-1. – S. 55-59

9. Popova O.B., Popov B.K., Klyuchko V.I. Poluchenie kornya binarnogo dereva sistemy voprosov i otvetov // Sovremennye problemy nauki i obrazovaniya. – 2013. №3. – S. 8

10. Popova O.B., Popov B.K., Klyuchko V.I. Problema sokrashcheniya vremeni vybora metodov upravleniya bolshimi sistemami (BS) // Sovremennye problemy nauki i obrazovaniya. – 2013. №1. – S. 163

11. Popova O.B., Popov B.K., Klyuchko V.I. Analiz svyazey v realnoy i tekhnicheskoy sistemakh protsessa optimizatsii // Mezhdunarodnyy zhurnal eksperimentalnogo obrazovaniya. – 2013. №10-2. – S. 405-408

12. Popova O.B., Popov B.K. Zamena realnoy sistemy (protsess vybora metoda optimizatsii) na tekhnicheskuyu sistemu (programma-sovetchik «Optimel») // Sovremennye problemy nauki i obrazovaniya. – 2012. – №5. – S. 132

13. Sistemnyy analiz protsessa vybora metoda optimizatsii informatsionnoy sistemy: monografiya / O.B. Popova, B.K. Popov, V.I. Klyuchko; FBGOU VPO «Kuban. gos. tekhnol. un-t». – Krasnodar: Izdatelskiy Dom – Yug, 2012 – 135 s.

14. Popova O.B., Popov B.K. Intellektualnaya informatsionnaya sistema vybora «Optimel». Patent na izobretenie RUS № 2564641 ot 27.05.2014.

15. Popova O.B. Uchastie protsessa optimizatsii v razvitii slozhnykh tekhnicheskikh sistem. Dep. v VINITI №257-V2010 07.05.2010.

*ENTROPY IN THE DECISION TREE, THE RECEIPT OF WHICH IS  
APPROACHED TO THE NATURAL INTELLIGENCE*

**O.B. POPOVA**

*Kuban State Technological University,  
2, Moskovskaya st., Krasnodar, Russian Federation, 350072,  
e-mail: popova\_ob@mail.ru*

The actual problem is the classification of a given knowledge domain, where they be submitted no encoded information in a certain way, but by way close to the natural intelligence, which structures the obtained knowledge on a different principle. Classify known answers to the questions, so that would solve the current problem. For understanding the knowledge in such way is suitable new method of forming a decision tree, close to the natural intelligence. The article considers how the entropy is associated with the emergence of

knowledge, classification of previous knowledge and the definitions used in the decision tree. The latter is necessary to compare the algorithm for constructing the decision tree, approximate to the natural intelligence with traditional methods. The dependence of the entropy by a subset of the properties of elements of the set. It is established that in order to reduce the entropy and the classification is necessary to define the properties for subsets, which gradually decreases the value. The condition for the termination of division to a subset of the number of elements is a subset of less than 2. The recursive repetition of simple action makes the algorithm effective and easy to implement a binary tree of questions and answers system creates a simple production rules.

**Key words:** entropy, the natural intelligence, knowledge, production rules, the decision support tree, a binary system of questions and answers tree.