ПРИМЕНЕНИЕ ДЕРЕВЬЕВ ПРИНЯТИЯ РЕШЕНИЙ ПРИ ОПРЕДЕЛЕНИИ ШАБЛОНОВ ДАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

Усов А.Е.[1], Варламов А.А.[2], Бабкин О.В.[3], Дос Е.В.[4], Мостовщиков Д.Н.[5]
Email: Usov1159@scientifictext.ru

[1]Усов Алексей Евгеньевич – ведущий архитектор;
[2]Варламов Александр Александрович – старший архитектор;
[3]Бабкин Олег Вячеславович – старший архитектор;
[4]Дос Евгений Владимирович – архитектор;
[5]Мостовщиков Дмитрий Николаевич – старший архитектор,
системный интегратор «Li9 Technology Solutions»,
г. Райли, Соединенные Штаты Америки

Аннотация: *рассмотрены методы разработки алгоритмов обнаружения знаний в базах данных как базового подхода выделения значимых образцов (шаблонов) в структуре больших наборов данных. В рамках разработанной методологии выделены две группы алгоритмов обнаружения знаний: кластеризация объектов, классы которых изначально не определены, и методы индуктивного обучения, в рамках которых на основе заданного набора классов определяется принадлежность к ним объекта исследования. Предложен оригинальный подход в области обнаружения знаний в базах данных, в основу которого положены методы классификации, что базируются на таком средстве поддержки принятия решений, как дерево принятия решений. Разработанная методика позволяет проводить анализ как на основе заданных шаблонов классификации данных, так и выделять новые признаки информационных объектов исследуемого набора и его классов, включая признаки высокого порядка, как, например, сходство между классами, характеристики классов и потенциальные ошибки представленного набора данных.*

Ключевые слова: *информационные системы, методы классификации, обнаружение знаний в базах данных, дерево принятия решений, C4.5, ID3, FTree.*

APPLICATION OF DECISION TREES AT DEFINING INFORMATION SYSTEM PATTERNS

Usov A.Ye.[1], Varlamov A.A.[2], Babkin O.V.[3], Dos E.V.[4], Mostovshchikov D.N.[5]

[1]Usov Aleksey Yevgenyevich – Lead Systems Architect;
[2]Varlamov Aleksandr Aleksandrovich – Senior Solution Architect;
[3]Babkin Oleg Vyacheslavovich – Senior System Architect;
[4]Dos Evgeniy Vladimirovich – System Architect;
[5]Mostovshchikov Dmitriy Nikolayevich – Senior System Architect,
IT INTEGRATOR «LI9 TECHNOLOGY SOLUTIONS»,
RALEIGH, UNITED STATES OF AMERICA

Abstract: *methods for the development of knowledge discovery algorithms in databases are considered as a basic approach of significant samples detection at big data sets. Within the framework of the developed methodology, two groups of knowledge detection algorithms are distinguished: clustering objects with undefined classes and methods of inductive learning for determined objects by given set of classes. An original approach in the field of knowledge discovery at databases is proposed, which is based on classification methods based on a decision support tool such as a decision tree. The developed technique allows analyzing both on the basis of predetermined data classification patterns and highlighting new features of information objects of the data set and its classes, including patterns of a higher order, such as the similarity between classes, characteristics of classes and potential errors of the presented data set.*

Keywords: *information systems, classification methods, databases' knowledge discovery, decision tree, C4.5, ID3, FTree.*

УДК 331.225.3

Введение

Методика обнаружения знаний в базах данных (KD: Knowledge Discovery) может быть определена как процесс выделения актуальных шаблонов при нейросетевом анализе больших наборов данных [1]. На сегодняшний день KD лежит в основе построения методов управления с прогнозирующими моделями, как наиболее продуктивного метода работы с базами данных, полученными в результате широкомасштабного исследования. В рамках данной работы, тем не менее, предлагается разделять задачи анализа данных и прогнозирования, и в соответствии с данным подходом строить математические

модели KD. Это позволит увеличить эффективность анализа через соотнесение исполняемого алгоритма и класса задачи, что обуславливает ***актуальность исследования*** проведенного в рамках данной работы.

Анализ последних исследований и публикаций в данной области показал приоритет использования при кластеризации алгоритмов обучения без учителя (unsupervised learning), которые группируют объекты предметной области (domain objects) по признаку их сходства [2], что соответствует парадигме индуктивного обучения. Кроме того были рассмотрены методы подразумевающие наличие предварительной информации об оптимальной стратегии кластеризации, например, заданное количество кластеров на основе которых нейросетевые алгоритмы определяют их центроиды и границы [3].

Наиболее широко используемым методом индуктивного обучения является средство поддержки принятия решений, известное как дерево принятия решений (decision tree). Данный подход можно разделить на две базовые группы:

- деревья классификации (classification trees), где прогнозируемым результатом работы алгоритма (например, ID3 [4] или его расширенная версия C4.5 [5], где выбор атрибута происходит на основании нормализованного прироста информации) является выделение классов данных;
- деревья регрессии, где прогнозируемый результат работы алгоритма (алгоритм CART [6]) может быть представлен в числовой форме.

При этом построение деревьев классификации рассматривается как более сложная и значимая задача индуктивного обучения [7]. Большая часть работ в этой области посвящена построению прогностических моделей, в то время как более важным является выделение ключевых атрибутов объектов набора данных [8, 9]. Процесс классификации может быть рассмотрен прохождение пути от корня дерева принятия решений к листьям [10], которое содержит значения атрибутов. В работе [11], тем не менее, предлагается данный подход предлагается расширить в рамках KD, таким образом, все узлы дерева помимо листьев содержат наборы примеров классов. Соответственно, прохождение от корня к узлу определяет уровень сходства классов. Предложенная методология анализа FTree (Filtered Tree) берет за основу алгоритм для роста дерева принятия решений (ID3, C4.5 или др.) с помощью которого производится анализ формы дерева. При этом остается возможность заложить в модель предварительный набор знаний.

Целью работы, таким образом, стала разработка методики получения точной модели предметной области на основе методов индуктивного обучения и деревьев принятия решений, которая производит анализ наибольшего числа объектов предметной области.

1. Методология построения дерева принятия решений при построении моделей мониторинга информационных систем

Дерево принятия решений в общем случае может быть представлен как ориентированный ациклический граф (DAG: Directed Acyclic Graph). Что касается структуры данного графа, следует отметить, что все его узлы, помимо корня имеют одно входящее ребро, а у корня, соответственно нет входящего ребра [1]. Аналогично, узлы без исходящих ребер называются листьями, а все остальные узлы — внутренними узлами. Эффективность алгоритма дерева принятия решений определяется через эффективность проведения классификации на основе данной структуры в отношении объектов, которые не входят в обучающий набор.

Пример построения алгоритма дерева принятия решений показан на рис. 1 и рис. 2. В качестве объекта моделирования была выбрана базовая схема системы выявления кибер-угроз (IDS: Intrusion Detection System) распределенной информационной системы (рис. 1).

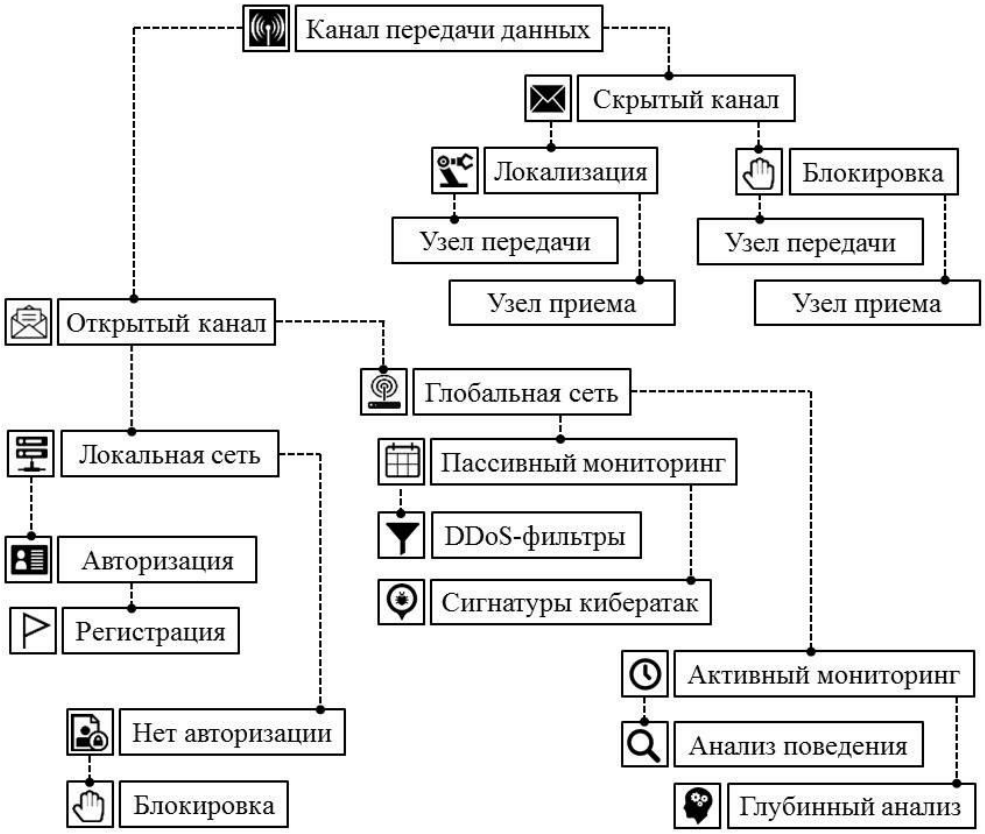

Рис. 1. Базовая схема системы выявления кибер-угроз распределенной информационной системы

На рис. 1 наглядно представлены фундаментальные основы работы комплекса IDS, схема включает в себя весь набор ключевых элементов, которые отвечают за мониторинг и блокировку потенциальных кибер-угроз инфраструктуры распределенной информационной системы, как внутренних, так и внешних. Тем не менее, такая схема не является интуитивно понятной для неспециалиста в данной области, например, научного сотрудника, который на основе математического моделирования хочет определить эффективность работы комплекса и подготовить рекомендации для его усовершенствования или масштабирования. Для решения поставленной задачи может быть предложено построить алгоритм дерева принятия решений (рис. 2).

Рис. 2. Алгоритм дерева принятия решений системы выявления кибер-угроз

Корнем дерева принятия решений, представленного на рис. 2, является «канал передачи данных» а листьями — «регистрация» и «блокировка». Как можно видеть, построение дерева принятия решений происходит через разделения базового набора элементов на подмножества атрибутов, которые далее рекурсивно разделяются на меньшие подмножества-домены до тех пор, пока анализ не позволит выйти на листья дерева принятия решений. Все доменные объекты должны быть представлены с помощью пар атрибут-значение (на уровне классификации или числовой оценки). Наиболее репрезентативный атрибут представленного алгоритма «скрытый канал», он позволяет провести однозначную классификацию программного кода как кибер-угрозы для информационного ресурса. Во всех остальных случаях, соответственно, подмножество должно быть разбито на меньшие подмножества. Такой подход в значительной мере упрощает схему IDS, но при этом он может быть положен в основу алгоритмов моделирования широкого спектра задач по оценке мониторинга и защиты распределенных информационных систем

2. Оптимизация методов математического моделирования алгоритмов дерева принятия решений

Как было показано в предыдущем разделе, ключевая задача, которая решается при построении алгоритмом деревьев принятия решений — это выбор оптимального атрибута для разделения узла. Эффективность решения данной задачи определяет эффективность дерева принятия решений, равно как и его структуру. На сегодняшний день в данной области широко используется универсальный подход [12], который базируется на оценке уровня беспримесности узла (NID: node impurity degree).

Так, например, соотнесение уровня беспримесности узла T_n по отношению к уровню беспримесности его дочерних узлов $T_{n,k} \in \{T_{n,1}; T_{n,K}\}$ относительно атрибута a_i, который для узла T_n, соответственно, может принимать K значений, что позволяет определить эффективность разбиения узла на подмножество, рассчитывается как:

$$\delta(a_i) = NID(T_n) - \sum_{k=1}^{K} \frac{NUM(T_{n,k}) \cdot NID(T_{n,k})}{NUM(T_n)}, \qquad (1)$$

где $NID()$ — функция оценки уровня беспримесности узла, а $NUM()$ определяет количество примеров, связанных с узлом.

Как можно видеть, в рамках данного подхода есть возможность построить алгоритмы деревьев регрессии на основе коэффициента Джини $G(T_n, C)$, который определяет разницу между распределением вероятности значений атрибутов $P_{n,c} \in \{P_{n,1}; P_{n,C}\}$.

$$\begin{cases} \delta(a_i) = G(T_n) - \sum_{k=1}^{K} \frac{NUM(T_{n,k}) \cdot G(T_{n,k})}{NUM(T_n)} \\ G(T_n) = 1 - \sum_{c=1}^{C} (P_{n,c})^2 \end{cases}, \quad (2)$$

где C, таким образом, определяет полный набор меток класса.

Аналогичным образом при построении деревьев классификации в качестве функции оценки уровня беспримесности узла при соотнесении уровня беспримесности отдельного узла по отношению к уровню беспримесности его дочерних узлов можно использовать показатель энтропии $G(T_n)$:

$$\begin{cases} \delta(a_i) = H(T_n) - \sum_{k=1}^{K} \frac{NUM(T_{n,k}) \cdot H(T_{n,k})}{NUM(T_n)} \\ H(T_n) = -\sum_{c=1}^{C} P_{n,c} \cdot log_2(P_{n,c}) \end{cases}. \quad (3)$$

Как можно видеть, в рамках данного определения $\delta(a_i)$ может быть описана, как прирост информации (information gain).

3. Разработка метода построения деревьев классификации на основе оценки уровня беспримесности узла и показателя энтропии

Развитие метода построения деревьев классификации на основе оценки уровня беспримесности узла и показателя энтропии возможно при подборе более релевантного атрибута. Так, например, подход может заключаться в соотнесении разбиения и атрибута, на основе которого было проведено данное разбиение. Оптимальным атрибутом является тот, который вызывает разбиение, соответствующее правильному разбиению из обучающего набора соответствующему этому узлу. При этом выбирается нормированная метрика на основе энтропии, определенная в множестве разбиений. Энтропия разбиения, таким образом, может быть определена как неопределенность соответствия случайно выбранного объекта определенному классу.

Рассмотрим разбиение, описываемое через $R \in \{R_1, \dots R_i, \dots R_I\}$ конечного множества элементов базы данных D. В таком случае энтропия $H()$ множества разбиений R, которое определяется через мощность $|R_i|$ (соответственно, мощность множества элементов базы данных равна $|D|$) может быть рассчитана на основе следующей системы уравнений:

$$\left[\begin{cases} H(R) = -\sum_{i=1}^{I} P_i \cdot log_2(P_i) \\ P_i \neq 0 \end{cases}, \quad (4)\right.$$
$$\left.\begin{cases} H(R) = 0 \\ P_i = 0 \end{cases}\right.$$

где:

$$P_i = \frac{|R_i|}{|D|}. \quad (5)$$

При этом расстояние между двумя разбиениями (например, $R \in \{R_1, \dots R_i, \dots R_I\}$ и $S \in \{S_1, \dots S_j, \dots S_J\}$ в данном случае определяется как:

$$\delta(R,S) = \frac{H(R|S) + H(S|R)}{H(R \cap S)}, \quad (6)$$

причем $H(R|S)$, $H(S|R)$ и $H(R \cap S)$ определяются через систему уравнений:

$$\begin{cases} H(R|S) = -\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{|R_i \cap S_i|}{|D|} \cdot log_2\left(\frac{|R_i \cap S_i|}{|S_i|}\right) \\ H(S|R) = -\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{|R_i \cap S_i|}{|D|} \cdot log_2\left(\frac{|R_i \cap S_i|}{P_i}\right) \\ H(R \cap S) = -\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{|R_i \cap S_i|}{|D|} \cdot log_2\left(\frac{|R_i \cap S_i|}{|D|}\right) \end{cases} \quad (7)$$

Данный математический аппарат может быть положен в основу построения дерева решений (рис. 3):

- работа с заданным набором базы знаний;
- работа с атрибутами, значения которых не заданы;
- масштабируемость структуры дерева знаний.

В области информационных технологий, и в частности в области построения алгоритмов мониторинга распределенных информационных систем существуют базы знаний, которые могут быть использованы при построении деревьев принятия решений. При этом в базовой схеме построения алгоритмов деревьев принятия решений не используются заданные набором базы знаний, поэтому подходы в могут разниться от использования данных наборов для определения атрибутов разбиения до их применения в построении модели организации структуры доменов.

Что касается работы с атрибутами, значения которых не заданы, то в данном случае также есть несколько подходов. Наиболее эффективным методом является рассмотрение статистической выборки и выделение наиболее часто встречающихся значений атрибута среди объектов, принадлежащих к соответствующему классу. В рамках данной работы также рассматривается подход, при котором таких атрибутов независимо от основного дерева принятия решений, с целью исключить ошибки при классификации. Это существенно увеличивает надежность и продуктивность средств мониторинга сетевых ресурсов распределенной информационной системы, но при этом является достаточно ресурсоемким подходом.

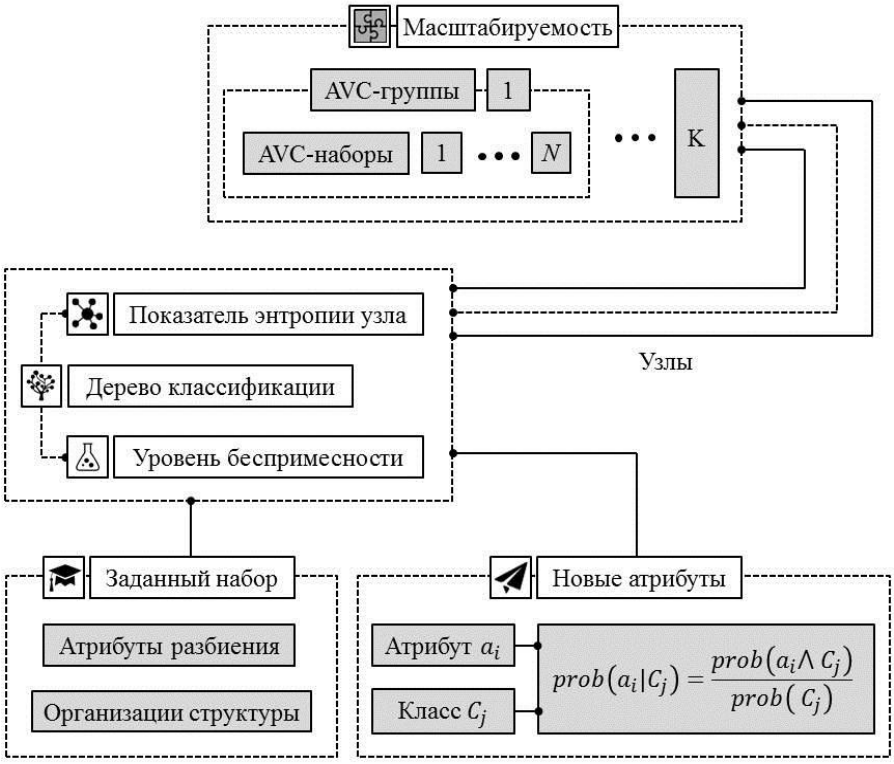

Рис. 3. Схема разработки деревьев классификации на основе оценки уровня беспримесности узла и показателя энтропии

Изначально на уровне математической модели подразумевалось, что структура дерева принятия решений должна быть фиксирована, в то время как в области анализа информационных систем в большинстве случаев возникает необходимость масштабирования данных алгоритмов в режиме реального времени. В данной работе предлагается использовать:

- методики SPRINT и CART [14], которые включают в себя создание списков атрибутов на этапе предварительной обработки всего множества информационных объектов при работе с деревьями регрессии;
- алгоритмы фреймворка RainForest, где используются как AVC-наборы, связывающие каждую пару атрибут-значение с меткой класса, так и AVC-группы, которые являются группами AVC-наборов, связанных с узлами дерева принятия решений [15].

Таким образом, была разработана комплексная методология построения алгоритмов деревьев принятия решений, которые могут быть поставлены за основу математического моделирования широкого спектра задач, связанных с мониторингом информационных систем.

Выводы

В результате проведенного анализа была разработана методология построения алгоритмов деревьев принятия решений в частности предложены:

1. обобщенная схема построения деревьев классификации;
2. математический аппарат классификации на основе оценки уровня беспримесности узла и показателя энтропии;
3. подходы, которые включают в себя работу с заданным набором базы знаний, работу с атрибутами, значения которых не заданы и масштабируемость структуры дерева знаний.

Было показано, что данная методология может быть использована при моделировании задач, связанных с анализом распределенных информационных систем.

Список литературы / References

1. *Maimon O. and Rokach L., editors.* Data Mining and Knowledge Discovery Handbook, 2nd ed. Springer, 2010.
2. *Luo J., Wu, Q. & Zhu L.,* 2013. Object-oriented full-time domain moving object data model. Journal of Computer Applications, 33(4), 1015-1017. doi:10.3724/sp.j.1087.2013.01015.

3. *Rajamohamed R. & Manokaran J.*, 2017. Improved credit card churn prediction based on rough clustering and supervised learning techniques. Cluster Computing, 21 (1), 65-77. doi:10.1007/s10586-017-0933-1.
4. *Zenghong W., Yufen C. & Jun Z.*, 2010. Adaptive rules mining in ACVis based on ID3 algorithm in decision tree. 2010 The 2nd Conference on Environmental Science and Information Application Technology. doi:10.1109/esiat.2010.5568899.
5. Symbology of the Logical Decision Tree, 2017. Decision-Making Management, 99-100. doi:10.1016/b978-0-12-811540-4.09979-8.
6. *Radoglou-Grammatikis P.I. & Sarigiannidis P.G.*, 2018. An Anomaly-Based Intrusion Detection System for the Smart Grid Based on CART Decision Tree. 2018 Global Information Infrastructure and Networking Symposium (GIIS). doi:10.1109/giis.2018.8635743.
7. *Pazzani M.J.* Knowledge discovery from data? IEEE Intelligent Systems, 15(2):10–13, 2000.
8. *Armengol E.* Building partial domain theories from explanations. Knowledge Intelligence, 2/08:19–24, 2008.
9. *Armengol E. and Plaza E.* Discovery of toxicological patterns with lazy learning. In V. Palade, R.J. Howlett and L. Jain, editors, KES-2003, number 2774 in Lecture Notes in Artificial Intelligence. Pages 919–926. Springer, 2003.
10. *Armengol E.* Usages of generalization in CBR. In R.O. Weber and M. M. Richter, editors, ICCBR-2007. Case-based Reasoning and Development, number 4626 in Lecture Notes in Artificial Intelligence, pages 31–45. Springer-Verlag, 2007.
11. *Armengol E., García-Cerdaña A. and Dellunde P.* Experiences Using Decision Trees for Knowledge Discovery. Springer International Publishing AG, 2017.
12. *Moulana M. & Hussain M.A.*, 2016. An Optimized Decision Trees Approach for Knowledge Discovery Using Orthogonal Radom Matrix Projection with Outlier Detection. International Journal of Database Theory and Application, 9 (3), 87-94. doi:10.14257/ijdta.2016.9.3.10.
13. *López R. de Mántaras.* A distance-based attribute selection measure for decision tree induction. Machine Learning, 6:81–92, 1991.
14. *Shafer J.C., Agrawal R. and Mehta M.* Sprint: A scalable parallel classifier for data mining. InVLDB, pages 544–555, 1996.
15. *Gehrke J., Ramakrishnan R. and Ganti V.* RainForest - a framework for fast decision tree construction of large datasets. Data Mining and Knowledge Discovery, 4(2/3):127–162, 2000.