УДК 004.056

Е.В. Зубков
(СибГУТИ; e-mail: evz.nsk@gmail.com)

ПРИМЕНЕНИЕ ЭНТРОПИЙНОГО ПОДХОДА В ЗАДАЧАХ ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Предлагается оригинальная методика кластеризации событий при оценке информационной безопасности. Данная методика использует энтропийный подход и позволяет формировать кластеры с заданными характеристиками.

Ключевые слова: энтропия, кластеризация, информационная безопасность.

E.V. Zubkov

APPLICATION OF ENTROPY APPROACH IN PROBLEMS OF ENSURING INFORMATION SECURITY

An original method clustering events during information security assessment is offered. This technique uses the entropy approach and allows forming clusters with predetermined characteristics.

Key words: data mining, entropy, clustering, information security events.

Статья поступила в редакцию Интернет-журнала 26 мая 2016 г.

Введение

Интенсивное развитие информационных технологий, наблюдаемое в последние десятилетия, создало предпосылки к накоплению и обработке информации, которая первоначально преимущественно использовалась в роли инструмента для изучения различных процессов и явлений. Усложнение структуры данных, увеличение их объёмов, появление новых форм данных привело к тому, что информация, независимо от своей предметной области, сама стала объектом исследований.

Совокупность методик, предназначенных для обработки больших массивов данных и извлечения из них дополнительной (скрытой) информации, эволюционно была сведена в единую предметную область и объединена общим термином – ***"методы интеллектуального анализа данных" (МИАД)***. В англоязычной литературе для её обозначения используют термин – "Data Mining". К основным задачам МИАД относят: классификацию, кластеризацию, прогнозирование, поиск ассоциативных правил и т.д. [1-3].

Несмотря на междисциплинарный характер МИАД, предметная область все же может накладывать существенные ограничения на возможность их применения. В данной статье исследуется статистика ***событий информационной безопасности (СИБ)***.

Источником таких событий является ***система обнаружения вторжений (СОВ)***, которая является важнейшей составляющей системы информационной безопасности. Одна из основных проблем, связанных с исследованием сетевых СИБ, заключается в их большом количестве. В связи с этим приобретает актуальность задача сокращения информационных сущностей, требующих экспертного анализа. Отметим, что признаки СИБ измерены в номинальной шкале и имеют высокие значения вариативности. Стандартные средства СУБД не всегда обеспечивают желаемый результат, поскольку количество результирующих элементов при группировке по всем значениям одного либо нескольких признаков будет сопоставимо с количеством исходных. Решение задачи лежит в плоскости применения МИАД для кластеризации СИБ.

Кластеризацией называют процесс распознавания внутренних правил объекта данных. Объекты группируют в форме классов связанных объектов, то есть кластеров в зависимости от выбранных метрик. Используемые метрики основаны на значениях свойств элементов. Различие между классификацией и кластеризацией состоит в том, что классификацию применяют для распределения элементов по заранее известным классам, а кластеризацию – для поиска неустановленных правил классификации в перемешанных наборах данных. Условно кластеризацию можно рассматривать как автоматическую классификацию.

Достаточно часто методики кластеризации основаны на вычислении расстояния между объектами. Однако, если, как в нашем случае, признаки объекта измерены в номинальной шкале, определить его величину весьма затруднительно. Это потребует использования некоторых допущений и искусственных преобразований. Полученный суррогат не может в полной мере описать действительные отношения между объектами, что приведет к упрощению исходной модели и снижению качества конечного результата.

Предлагаемый подход выгодно отличается от подобных методик, поскольку изначально предназначен для работы с номинальными значениями. Заложенные в него принципы позволяют формировать группы однородных элементов за счет непосредственного анализа значений независимых признаков.

Общие принципы методики

В основу методики положены принципы, изложенные в [4]. В качестве критериев кластеризации используются следующие пороговые величины:

- минимальное количество элементов в кластере;

- минимальное значение однородности в кластере.

Основные этапы процесса представлены на рис. 1 и включают в себя:

• вычисление ***наиболее информативного признака (НИП)*** для исследуемого множества элементов;

• вычисление наиболее ***информативного значения (НИЗ)*** среди всех возможных значений НИП;

• выборку элементов из исходного множества элементов, у которых значение НИП соответствует НИЗ;

• вычисление значения однородности для полученных множеств;

• контроль однородности: если вычисленное значение меньше порогового значения, то полученное множество помещают в список исходных множеств и для него повторяют п.п. 1-5;

• контроль количества элементов: если количество элементов в множестве меньше порогового значения, данные элементы считают статистически малозначимыми и помещают в специальное множество для некластеризованных данных.

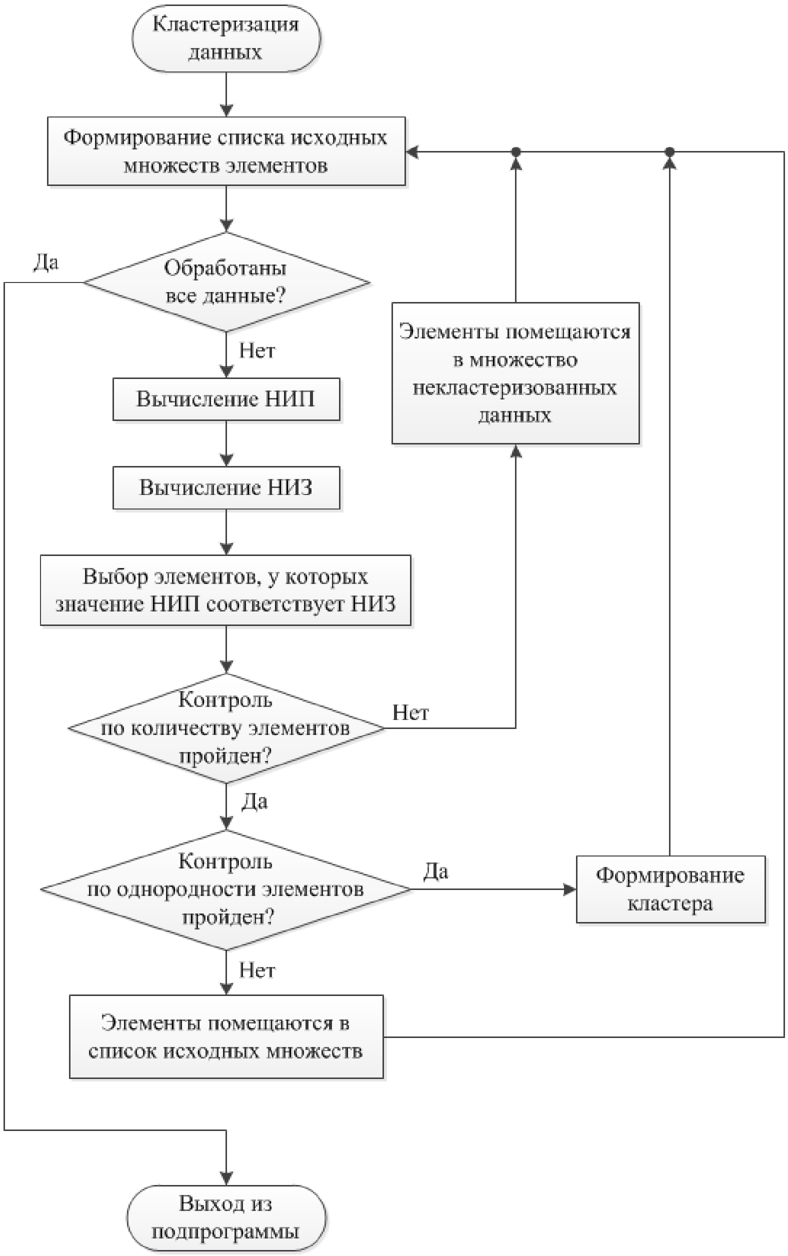

Рис. 1. Блок-схема алгоритма кластеризации данных

Группы элементов, удовлетворяющие обоим пороговым значениям, оформляются в виде кластеров. Для них формируется шаблон, который в последующем сохраняется в БД.

Определение наиболее информативного признака

Рассмотрим множество элементов $A{:}\,\{a_1, a_2, \dots, a_n\}$, где n – количество элементов в множестве. Каждый элемент a характеризуют набором из m признаков:

$$a \to \{v_1, v_2, \dots, v_m\}.$$

Каждый признак v_i является номинальной величиной и может принимать одно из предопределенных значений:

$$v_{i(r)} \in V_i{:}\,\{v_{i(1)}, v_{i(2)}, \dots, v_{i(s_i)}\},$$

где i – номер признака;

r – номер значения признака;

s_i – количество уникальных значений признака v_i, которые встречаются во множестве элементов A;

V_i – множество уникальных значений признака v_i;

$v_{i(r)}$ – значение r признака i.

Вероятность появления значения $v_{i(r)}$в произвольном элементе составляет:

$$p_{i(r)} = \frac{n_{i(r)}}{n},$$

где $n_{i(r)}$ – количество элементов, у которых признак v_i имеет значение $v_{i(r)}$.

Отметим, что для всех i верны утверждения:

$$\sum_{r=1}^{s_i} n_{i(r)} = n,$$
$$\sum_{r=1}^{s_i} p_{i(r)} = 1.$$

Энтропию каждого признака можно представить выражением следующего вида:

$$H_i = -\sum_{r=1}^{s_i} p_{i(r)} \log_2 p_{i(r)}.$$

Значение энтропии H_i будет находиться в некотором диапазоне $0 \le H_i \le H_{max_i}$, причем $H_i = 0$ соответствует ситуации с нулевой дисперсией, когда признак v_i у всех элементов имеет одинаковое значение. Максимального значения ($H_i = H_{max_i}$) энтропия достигает в случае с максимальной дисперсией, то есть когда каждое значение признака встречается равное количество раз.

Далее необходимо определить взаимное влияние признаков. Для каждой пары признаков v_i и v_j строим двумерную переменную, представленную в табл. 1, где МР – маргинальное распределение.

Таблица 1

Значения двумерной номинальной переменной

Значения признака v_i	**Значения признака v_j**				**МР v_j**
	$v_{j(1)}$	$v_{j(2)}$	…	$v_{j(s_j)}$	
$v_{i(1)}$	$n_{i(1),j(1)}$	$n_{i(1),j(2)}$	…	$n_{i(1),j(s_j)}$	$n_{i(1)}$
$v_{i(2)}$	$n_{i(2),j(1)}$	$n_{i(2),j(2)}$	…	$n_{i(2),j(s_j)}$	$n_{i(2)}$
…	…	…	…	…	…
$v_{i(s_i)}$	$n_{i(s_i),j(1)}$	$n_{i(s_i),j(2)}$	…	$n_{i(s),j(s_j)}$	$n_{i(s_i)}$
МР v_i	$n_{j(1)}$	$n_{j(2)}$	…	$n_{j(s_j)}$	n

Вероятность появления в элементе комбинации значений $v_{i(r)}$и $v_{j(q)}$ будем рассчитывать по формуле:

$$p_{i(r),j(q)} = \frac{n_{i(r),j(q)}}{n},$$

где $n_{i(r),j(q)}$ – количество элементов, у которых признаки v_i и v_jимеют значения $v_{i(r)}$и $v_{j(q)}$.

Соответственно энтропию для пары этих признаков можно вычислить следующим образом:

$$H_{ij} = -\sum_{r=1}^{s_i} \sum_{q=1}^{s_j} p_{i(r),j(q)} \, log_2 \, p_{i(r),j(q)}.$$

Очевидно, что H_{ij} может принимать значения на интервале $[0; H_i + H_j]$. В случае полной связи, то есть когда $s_i = s_j$ и когда каждому признаку $v_{i(r)}$ соответствует строго определенный признак $v_{j(q)}$, имеем

$$H_{ij} = H_i = H_j. \qquad (1)$$

При статистической независимости, когда связь между признаками отсутствует, получаем:

$$H_{ij} = H_i + H_j. \qquad (2)$$

Условная энтропия ($H_{j;i}$) показывает, какая часть энтропии остаётся, если становится известно значение признака v_i.

$$H_{j;i} = H_{ij} - H_i.$$

В случае, когда выполняется условие (1), условная энтропия равна нулю. Это означает, что вся информация о значениях признака v_j содержится в признаке v_i. При независимости признаков, то есть если выполняется условие (2), условная энтропия $H_{j;i}$ будет равна всей энтропии H_j признака v_j. Оценить часть энтропии признака v_j, которая будет объясняться значением признака v_i, позволяет следующее выражение:

$$H_i - H_{j;i} = H_i + H_j - H_{ij} = H_i - H_{i;j} = h_{ij}.$$

Наглядно эту зависимость удобно проследить с использованием диаграммы Венна (рис. 2).

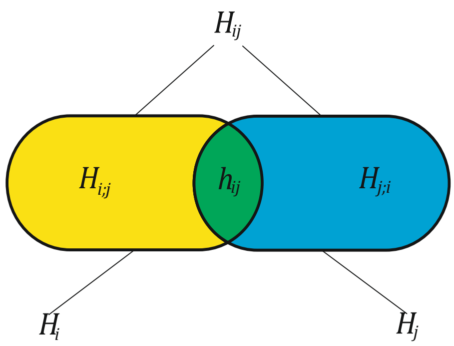

Рис. 2. Диаграмма Венна – зависимость между значениями для двумерной номинальной переменной

Относительный частный коэффициент влияния v_i на v_j можно тогда определить как отношение

$$\frac{h_{ij}}{H_i}.$$

Аналогично, частный коэффициент влияния v_j на v_i будет соответствовать

$$\frac{h_{ij}}{H_j}.$$

Указанный коэффициент считают хорошим показателем взаимного влияния признаков. Однако, как отмечают в [8], значение h_{ij} симметрично относительно v_i и v_j и лежит в диапазоне от 0 до $H_{ij} = H_i = H_j$. Поэтому более удобным и информативным показателем будет

$$\frac{h_{ij}}{H_{ij}},$$

который изменяется от 0 (в случае независимости признаков) до 1 (в случае полной связи). Рассчитаем средневзвешенное значение этого параметра для признака v_i относительно всех прочих признаков, иными словами для всех $j \neq i$, по формуле:

$$M\left\{\frac{h_{ij}}{H_{ij}}\right\} = \frac{\sum_{j\neq i}\frac{h_{ij}}{H_{ij}}H_{ij}}{\sum_{j\neq i}H_{ij}} = \frac{\sum_{j\neq i}h_{ij}}{\sum_{j\neq i}H_{ij}} = I_i.$$

Полученная величина I_i может служить мерой информативности признака v_i относительно остальных признаков. Чем выше значение I_i, тем больший объём совокупной информации несёт в себе признак v_i о значениях признаков $v_{j\neq i}$. Признак, обладающий наибольшим значением I_i, назовем наиболее информативным признаком (НИП).

Определение наиболее информативного значения признака

Очевидно, что в общую информативность признака вносит свой вклад каждое его значение. Допустим, что некоторое значение НИП встречается с большим числом значений другого признака, которое, в свою очередь, также встречается с различными значениями НИП. Степень зависимости здесь будет проявляться достаточно слабо. Второе значение НИП, напротив, соотносится с вполне ограниченным количеством значений другого признака, которые, в свою очередь встречаются, в основном (или только) с этим значением НИП. Справедливо считать, что второе значение более информативно. Значение, у которого описанное качество относительно прочих признаков выражено в наибольшей степени, будем называть наиболее информативным значением (НИЗ).

Рассмотрим множество элементов в трех проекциях. Случай первый: выделим в исходном множестве подмножество элементов, у которых признак v_i принимает значение $v_{i(r)}$, тогда для всех остальных элементов будет выполняться условие $v_i \neq v_{i(r)}$ (рис. 3).

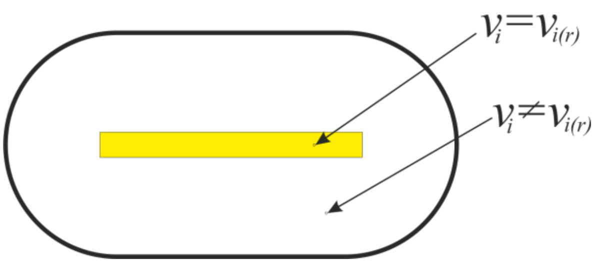

Рис. 3. Распределение элементов – случай первый

Для этого случая энтропию можно вычислить по формуле:

$$H_{i(r)} = -\left(p_{i(r)}\,log_2\,p_{i(r)} + p_{\overline{i(r)}}log_2 p_{\overline{i(r)}}\right), \tag{3}$$

где $p_{i(r)}$ – вероятность появления элемента со значением $v_{i(r)}$ в признаке v_i, то есть $v_i = v_{i(r)}$;

$p_{\overline{i(r)}}$ – вероятность появления элемента со значением признака v_i, отличным от значения $v_{i(r)}$, то есть $v_i \neq v_{i(r)}$.

Далее определим энтропию значений признака v_j $(j \neq i)$ относительно значения $v_{i(r)}$. Очевидно, существуют такие значения признака v_j, которые хотя бы один раз встречаются в комбинации с $v_{i(r)}$. Для определенности, обозначим множество таких значений признакаv_j, как $V_{j,i(r)}$. Выделим элементы, у которых $v_j \in V_{j,i(r)}$ в отдельное множество. Остальные элементы составят второе множество, где значение признака v_j ни разу не встречается в комбинации с$v_{i(r)}$, то есть $v_j \notin V_{j,i(r)}$ (рис. 4).

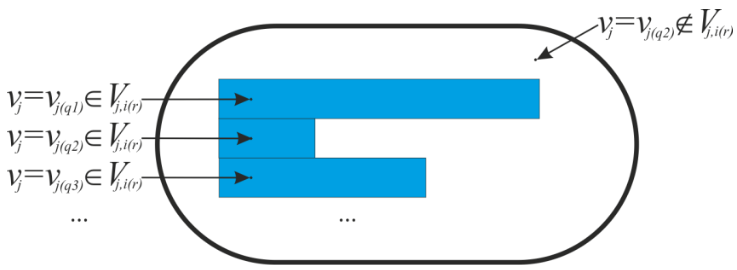

Рис. 4. Распределение элементов – случай второй

Энтропию вычисляют по формуле:

$$H_j^{i(r)} = -\left(p_{i(r),j}log_2 p_{i(r),j} + p_{\overline{i(r)},j}log_2 p_{\overline{i(r)},j}\right), \tag{4}$$

где $p_{i(r),j}$ – вероятность появления элемента с любым значением в признаке v_j, которое хотя бы раз встречается в комбинации со значением $v_{i(r)}$ в признаке v_i;

$p_{\overline{i(r)},j}$ – вероятность появления элемента с любым значением в признаке v_j, которое ни разу не встречается в комбинации со значением $v_{i(r)}$ в признаке v_i.

Теперь необходимо рассчитать энтропию комбинаций значения $v_{i(r)}$ со значениями признака v_j. В произвольном элементе $a \in A$ каждое значение $v_j = v_{j(q)}$ может находиться в одном из трёх состояний относительно $v_{i(r)}$ (рис. 5):

- $v_j = v_{j(q)}$ и $v_i = v_{i(r)}$;
- $v_j = v_{j(q)}$ и $v_i \neq v_{i(r)}$, но существуют такие элементы $a \in A$, где выполняется условие: $v_j = v_{j(q)}$ и $v_i = v_{i(r)}$, то есть $v_{j(q)} \in V_{j,i(r)}$;
- $v_j = v_{j(q)}$ ни разу не встречается в комбинации с $v_{i(r)}$, то есть $v_{j(q)} \notin V_{j,i(r)}$.

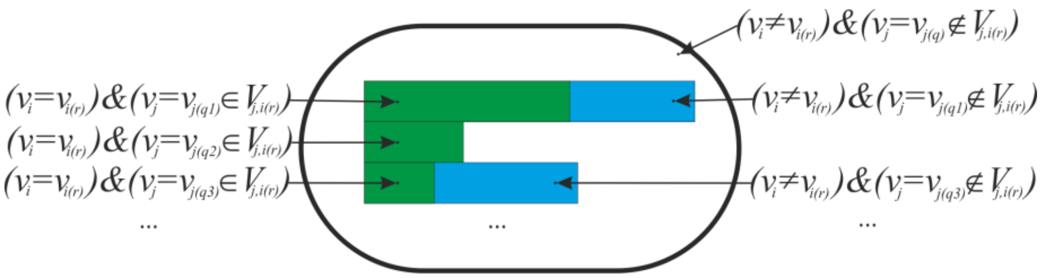

Рис. 5. Распределение элементов – случай третий

Энтропию такого распределения можно выразить следующим образом:

$$H_{ij}^{i(r)} = -\left(\sum_{q=1}^{s_j^{i(r)}}\left(p_{i(r),j(q)}\, log_2\, p_{i(r),j(q)}\right) + \sum_{q=1}^{s_j^{i(r)}}\left(p_{\overline{i(r)},j(q)}\, log_2\, p_{\overline{i(r)},j(q)}\right) + p_{\overline{i(r)},j}\, log_2 p_{\overline{i(r)},j}\right), \qquad (5)$$

где $p_{i(r),j(q)}$ – вероятность появления элемента со значением $v_{j(q)}$ в признаке v_j, которое хотя бы раз встречается в комбинации со значением $v_{i(r)}$ в признаке v_i, то есть выполняются условия: $v_j = v_{j(q)} \in V_{j,i(r)}$ и $v_i = v_{i(r)}$;

$p_{\overline{i(r)},j(q)}$ – вероятность появления элемента со значением $v_{j(q)}$ в признаке v_j и значением признака v_i, отличным от значения $v_{i(r)}$, то есть выполняются условия: $v_j = v_{j(q)} \in V_{j,i(r)}$ и $v_i \neq v_{i(r)}$;

$p_{\overline{i(r)},j}$ – вероятность появления элемента с любым значением в признаке v_j, которое ни разу не встречается в комбинации со значением $v_{i(r)}$ в признаке v_i, то есть выполняется условие: $v_j \notin V_{j,i(r)}$;

$s_j^{i(r)}$ – количество комбинаций значения $v_{i(r)}$ признака v_i со значениями признака v_j.

Обобщая сказанное, видим, что в первом случае мы располагаем информацией только о значении признака $v_i = v_{i(r)}$. Во втором случае – информацией о значениях признака v_j,которые встречаются в сочетании с $v_{i(r)}$. Третий случай – консолидирующий. Доступна полная информация о сочетаниях ния $v_{i(r)}$с различными значениями признака v_j.

Вычисленные значения энтропии (3-5) можно соотнести между собой следующим образом:

$$H_{ij}^{i(r)} < H_{i(r)} + H_j^{i(r)},$$

если значение $v_{i(r)}$несет в себе некоторое количество информации о значениях признака $v_j \in V_{j,i(r)}$:

$$H_{ij}^{i(r)} = H_{i(r)} + H_j^{i(r)},$$

если значение $v_{i(r)}$не содержит такой информации.

Выражение

$$h_{ij}^{i(r)} = H_{i(r)} + H_j^{i(r)} - H_{ij}^{i(r)}$$

позволяет определить часть энтропии значений признака $v_j \in V_{j,i(r)}$, которая становится известна, если известно значение $v_{i(r)}$.Заметим, что поскольку $v_{i(r)}$ не комбинирует со значениями $v_j \notin V_{j,i(r)}$, то есть вероятность появления элемента у которого $v_i = v_{i(r)}$ и $v_j = v_{j(q)} \notin V_{j,i(r)}$ равна 0, то справедливо будет считать, что $h_{ij}^{i(r)}$ отражает информационную значимость $v_{i(r)}$ относительно всех значений признака v_j.

Тогда относительный коэффициент влияния $v_{i(r)}$на v_j будет

$$\frac{h_{ij}^{i(r)}}{H_{ij}^{i(r)}}.$$

Теперь можно рассчитать средневзвешенное значение этого параметра для всех $j \neq i$ относительно $v_{i(r)}$:

$$M\left\{\frac{h_{ij}^{i(r)}}{H_{ij}^{i(r)}}\right\} = \frac{\sum_{j\neq i}\frac{h_{ij}^{i(r)}}{H_{ij}^{i(r)}}H_{ij}^{i(r)}}{\sum_{j\neq i}H_{ij}^{i(r)}} = \frac{\sum_{j\neq i}h_{ij}^{i(r)}}{\sum_{j\neq i}H_{ij}^{i(r)}} = I_{i(r)}.$$

Вычисленную величину можно считать мерой информативности ния $v_{i(r)}$относительно всех прочих признаков для текущей выборки элементов. Его значение может меняться от 0 до 1. Значение, обладающее максимальной информативностью, считаем НИЗ.

Контроль однородности выделенных кластеров

Полученные значения позволяют из общего массива данных сформировать кластер. НИП элементов этого кластера имеет значение соответствующее НИЗ. Таким образом, получаем два множества. В отношении каждого из них можно либо рекурсивно выполнить описанную выше процедуру, либо сохранить в БД в качестве кластера. Для принятия решения оценивают два параметра: количество элементов и однородность. Если однородность элементов множества выше предопределенного порогового значения, то это множество сохраняют в виде кластера, в противном случае выполняют контроль количества элементов. Если количество элементов превышает пороговое значение, то множество подвергают повторной обработке. Элементы, которые не удается объединить с соблюдением требований к однородности и минимальному количеству элементов в кластере, помещают в отдельное множество для некластеризованных данных.

Однородность является мерой степени "похожести" элементов в кластере и вычисляется по формуле:

$$u = \frac{\sum_{j=1}^{m} n_{max(j)}}{nm},$$

где n – количество элементов в исследуемом множестве;

m – количество признаков элемента;

$n_{max(j)}$– максимальное количество элементов с одинаковым значением признака v_j.

Диапазон значений однородности, таким образом, меняется от $\frac{1}{n}$, если значения всех признаков различны, до 1, если значения всех признаков одинаковые.

Выводы

В статье предложена оригинальная методика кластеризации сетевых СИБ, регистрируемых СОВ. В её основе лежит энтропийный анализ, который является составляющей компонентой МИАД. Ключевая особенность МИАД заключается в возможности обрабатывать данные с малым числом признаков, измеренных в номинальной шкале и обладающих высоким значением вариативности. Путем последовательного вычисления НИП и НИЗ выполняется формирование кластеров с заданными характеристиками. В результате данные приобретают более структурированную форму, появляется дополнительная информация статистического характера – количество элементов в кластере и их однородность, все это упрощает работу с сетевыми СИБ.

Литература

1. ***Интуит.*** Национальный открытый университет. Лекция 4: Задачи Data Mining. Информация и знания. http://www.intuit.ru/studies/courses/6/6/lecture/164.

2. ***Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.*** Технологии анализа данных: DataMining, VisualMining, TextMining, OLAP. 2-е изд. СПб.: БХВ-Петербург, 2007. 384 с.

3. ***Harshna, Navneet K.*** Fuzzy Data Mining Based Intrusion Detection System Using Genetic Algorithm. January 2014. http://www.ijarcce.com/upload/2014/january/IJARCCE3I__a_harshna_fuzzy.pdf.

4. ***Мёллер Ф.*** Роль энтропии в номинальной классификации // Математика в социологии. Моделирование и обработка информации. М.: Мир, 1977. 385 с.