ИСПОЛЬЗОВАНИЕ КЛАСТЕРНОГО АНАЛИЗА ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОБЪЕДИНЕНИЯ ВЗАИМОСВЯЗЕЙ МЕЖДУ СИГНАЛАМИ

Дудко Я. В.

Дудко Ярослав Владимирович / Dudko Yaroslav Vladimirovich – аспирант,
кафедра информационных систем и технологий, факультет фундаментальной и прикладной информатики,
Юго-Западный государственный университет, г. Курск

Аннотация: *рассмотрен общий подход к решению задач объединения выявленных взаимосвязей между сигналами, описываемых в виде прецедентов, возникающих в распределенной управляющей системе; предложен метод на основе математического аппарата кластерного анализа, предназначенный для выявления устойчивых взаимосвязей между прецедентами и локализации источника возникновения нештатных ситуаций; в качестве основы для хранения выявленных взаимосвязей предложено использование механизма дерева решений с неограниченным количеством ветвей узла, построенного при помощи алгоритма C4.5.*

Ключевые слова: *прецедент, интеллектуальный анализ данных, надежность, метод, кластер, дерево решений, градиент.*

Восстановление состояния системы после обнаружения ее нештатного функционирования наряду с определением критериев достоверности данных информационного обмена играет важную роль при решении задачи обеспечения надежности в распределенных управляющих системах.

Отдельные обнаруживаемые прецеденты, появляющиеся в системе, целесообразно выделять в группы с устойчивыми взаимосвязями, используя методы интеллектуального анализа информационных потоков (Data Mining). Задача кластеризации успешно осуществляет подобные объединения [1].

На основе данных (свойств), описывающих сущность объектов (наблюдений, событий), можно произвести группировку этих объектов, что и является целью кластеризации. В пределах одного кластера объекты должны быть максимально похожими друг на друга. При этом они так же должны максимально отличаться от объектов из других кластеров. Задача кластеризации осуществляется тем эффективней и точнее, чем больше наблюдается сходств между объектами внутри кластера и различий между кластерами.

Благодаря отсутствию накладываемых ограничений на представление исследуемых объектов возможен анализ показателей разного рода: интервальные данные, частоты, бинарные данные и т.д. Для этого необходимым является измерение и сравнение переменных в нормализованном представлении.

Кластерный анализ позволяет сокращать размерность анализируемых данных и представлять их в наглядном структурированном виде.

Кроме того, кластерный анализ применяется к совокупностям временных рядов. При этом выделяются периоды схожести некоторых показателей и определяются группы временных рядов со схожей динамикой [2]. Для кластерного анализа можно выделить следующие группы задач:

1) задача классификации или разработки типологии;
2) задача анализа принципов группирования объектов;
3) задача формирования гипотез на основе исследования полученных данных;
4) задача проверки гипотез для определения входимости выделенных типов в имеющихся данных.

Как правило, при использовании кластерного анализа решаются одновременно несколько поставленных задач.

Следующие математические характеристики описывают кластер: размер кластера, его радиус, центр кластера и среднеквадратичное отклонение.

Центром кластера является среднее геометрическое место точек в пространстве переменных. Радиус кластера – это максимальное расстояние точек до центра кластера.

Размер кластера равен либо радиусу кластера, либо среднеквадратичному отклонению объектов данного кластера [3]. Объект входит в состав кластера, если его расстояние до центра кластера меньше радиуса кластера.

Кластерный анализ возможен при выполнении следующих условий [5]:

– рассматриваемая совокупность объектов может быть разбита на кластеры на основании признаков этих объектов;

– для сопоставления признаков выбраны правильные единицы измерения признаков (произведена их нормализация).

В данной статье предложен способ группирования выявленных взаимосвязей между сигналами при помощи кластерного анализа. Исходными данными является совокупность прецедентов, формируемая при помощи алгоритма поиска взаимосвязей между сигналами для определения нештатного функционирования систем [4].

Под кластером в данном методе будет пониматься группа прецедентов, выявленных на основе информации, хранящейся в базе знаний в виде временных рядов, и содержащих сведения о взаимосвязях между сигналами. Для вхождения прецедента в кластер необходимо наличие в его составе сигнала, присутствующего хотя бы в одном из прецедентов кластера. Близость прецедента к центру кластера определяется градиентом частоты возникновения прецедентов кластера. Прецедент, не входящий ни в один из имеющихся кластеров, образует новый кластер и является его центром.

Для кластеризации каждого из прецедентов выделены следующие этапы:

1. Определение кластера для прецедента

Для всех сигналов прецедента осуществляется поиск их вхождений в прецеденты из состава имеющихся кластеров. В случае выявления такого вхождения прецедент включается в состав кластера. Прецеденты, не имеющие общих сигналов ни с одним из кластеров, образуют новые кластеры.

2. Определение нового центра кластера, в состав которого был включен прецедент, с последующим пересчётом расстояний до центра кластера каждого прецедента.

Значение градиента частот [6] возникновения прецедентов определяет центр кластера:

$$\overrightarrow{grad F_{cl}} = (\frac{\partial F_{cl}}{\partial x_1};...;\frac{\partial F_{cl}}{\partial x_n}), \quad (1)$$

где $x_1 ... x_n$ – значения, обратные частотам возникновения прецедентов; F_{cl} – суммарное условие возникновения прецедентов в кластере: $F_{cl} = \Sigma F_i$.

Расстояние до центра кластера i-го прецедента вычисляется по формуле:

$$\left|\overrightarrow{grad F_{cl}}\right|_i = \left(\frac{\partial F_{cl}}{\partial x_1}\right)_i + ... + \left(\frac{\partial F_{cl}}{\partial x_n}\right)_i. \quad (2)$$

3. Построение неориентированного невзвешенного графа прецедентов в рамках кластера.

Прецеденты из состава кластера являются вершинами графа, а наличие общих сигналов в составе прецедентов определяют связи между ними. На рисунке 1 представлен пример построения кластера.

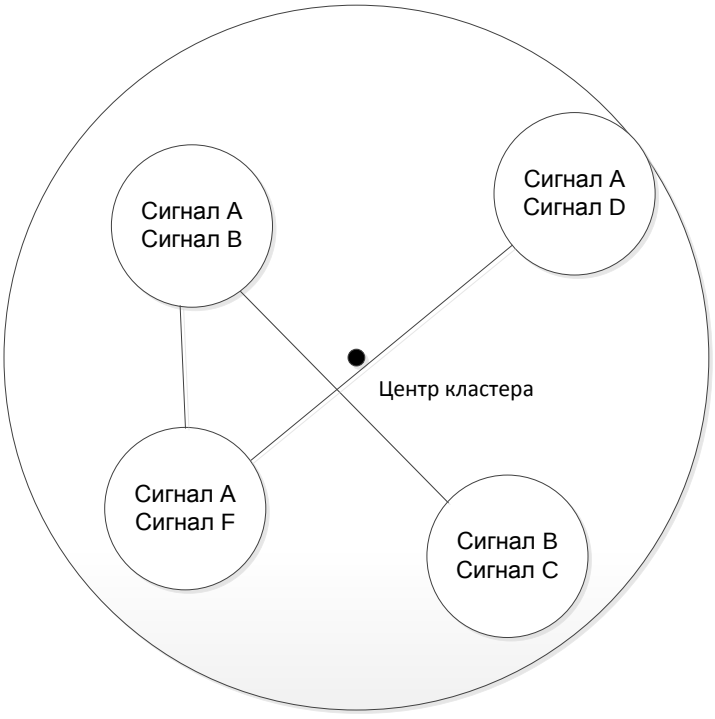

Рис. 1. Представление кластера прецедентов в виде графа

4. Осуществление выбора оптимального правила в графе кластера для выделения в его составе набора прецедентов, имеющих наибольшую взаимосвязанность, на основе классификационного алгоритма C4.5.

Для построения дерева решений с неограниченным количеством ветвей узла используется алгоритм C4.5 [7]. Данный алгоритм предназначен для решения исключительно классификационных задач, так как область его применения исключает все атрибуты кроме зависимых дискретных.

Кластерный анализ позволяет распределить выявленные прецеденты в кластеры на основе их взаимосвязей и частот возникновения, что делает возможным определение наиболее взаимосвязанных прецедентов для локализации источника возникновения нештатных ситуаций в распределенных управляющих системах.

Литература

1. *Дюк В., Самойленко А.* Data mining. Учебный курс. СПб.: Питер, 2001. 368 с.
2. *Чубукова И. А.* Data mining. М.: Бином, 2008. 384 с.
3. Технологии анализа данных: DataMining, VisualMining, TextMining, OLAP / Барсегян, Куприянов, Степаненко, Холод, Под ред. Барсегяна А. А. 2 изд. СПб.: БХВ-Петербург, 2007. 336 с.
4. *Башмаков А. И., Дудко Я. В.* Алгоритм обнаружения и анализа нештатных ситуаций // Информатика, вычислительная техника и управление. Ижевск: Системная инженерия. Научно-теоретический журнал, 2015. С. 100-104.
5. *Гитис Л. Х.* Кластерный анализ в задачах классификации, оптимизации и прогнозирования. М.: МГГУ, 2001. 103 с.
6. *Дубровин Б. А., Новиков С. П., Фоменко А. Т.* Современная геометрия методы и приложения: учебное пособие для физико-математических специальностей университетов. М.: Наука, 1986. 759 с.
7. *Hand D., Mannila H. and Smyth P.*, 2001. Principles of Data Mining. London: MIT Press. Pp: 197-201.