Инженерный вестник Дона, №3 (2016)
ivdon.ru/ru/magazine/archive/n3y2016/3695

Метод объединения выявленных взаимосвязей между сигналами с помощью кластерного анализа

А.И. Башмаков, Д.Р. Гуляева, Я.В. Дудко

Юго-Западный государственный университет, Курск

Аннотация: Предложен метод объединения выявленных взаимосвязей между сигналами, представленных в виде прецедентов распределенной управляющей системы, с помощью кластерного анализа для последующего выявления наиболее взаимосвязанных прецедентов с целью локализации источника возникновения нештатных ситуаций.
Ключевые слова: прецедент, кластер, дерево решений, градиент, интеллектуальный анализ данных, надежность, метод.

При решении задачи обеспечения надежности в распределенных управляющих системах наряду с определением критериев достоверности данных информационного обмена важную роль играет восстановление системы после обнаружения ее нештатного функционирования. При использовании методов интеллектуального анализа (Data Mining) информационных потоков целесообразно выделять отдельные обнаруживаемые прецеденты в группы с устойчивыми взаимосвязями. Осуществление подобных объединений может быть решено при помощи задачи кластеризации [1].

Целью кластеризации является группировка объектов (наблюдений, событий) на основе данных (свойств), описывающих сущность объектов. Объекты внутри кластера должны быть похожими друг на друга и отличаться от объектов, вошедших в другие кластеры. Чем больше похожи объекты внутри кластера и чем больше отличий между кластерами, тем точнее кластеризация. С помощью кластеризации средства Data Mining самостоятельно выделяют различные однородные группы данных.

Отсутствие накладываемых ограничений на представление исследуемых объектов позволяет анализировать разнородные показатели (интервальные данные, частоты, бинарные данные). Важным условием

является необходимость измерения и сравнения переменных в нормализованном представлении.

Кластерный анализ предназначен для сокращения размерности анализируемых данных и их представления в наглядном структурированном виде.

Кластерный анализ может применяться к совокупностям временных рядов, могут выделяться периоды схожести некоторых показателей и определяться группы временных рядов со схожей динамикой [2]. Выделяют следующие группы задач кластерного анализа:

1. задача разработки типологии или классификации;
2. задача исследования концептуальных схем группирования объектов;
3. задача выдвижения гипотез на основе исследования данных;
4. задача подтверждения гипотез для определения достоверности входимости выделенных типов в имеющихся данных.

В общем случае, при применении кластерного анализа решаются одновременно несколько указанных задач.

Кластер описывается следующими математическими характеристиками: центр, радиус, среднеквадратичное отклонение, размер кластера. Центр кластера – среднее геометрическое место точек в пространстве переменных. Радиус кластера – максимальное расстояние точек от центра кластера. Возможно возникновение ситуации, при которой невозможно определить принадлежность объекта к конкретному кластеру, используя математические процедуры.

Размер кластера определяется либо радиусом кластера, либо среднеквадратичным отклонением объектов для данного кластера [3]. Объект принадлежит кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера.

Процесс кластерного анализа допускает два предположения [7]:

– рассматриваемые признаки объекта допускают требуемое разбиение совокупности объектов на кластеры;

– при сопоставлении признаков были выбраны правильные масштабы или единицы измерения признаков (произведена их нормализация).

В данной статье предложен метод объединения выявленных взаимосвязей между сигналами с помощью кластерного анализа. Исходными данными является набор прецедентов, формируемый на основе алгоритма поиска взаимосвязей между сигналами для определения нештатного функционирования систем [4].

В контексте предложенного метода под кластером будет пониматься группа выявленных прецедентов, содержащих сведения о взаимосвязях между сигналами на основе информации, хранящейся в базе знаний в виде временных рядов [5]. Условием вхождения в кластер для прецедента является наличие в его составе сигнала, присутствующего хотя бы в одном из прецедентов кластера. Критерием близости прецедента к центру кластера [6] является значение градиента частоты возникновения прецедентов кластера. Если прецедент не входит ни в один из имеющихся кластеров, то он образует новый кластер и является его центром.

В представленном методе для кластеризации каждого из прецедентов определены следующие этапы:

1. Включение прецедента в состав одного из ранее выявленных кластеров либо создание для него нового кластера.

Для каждого из сигналов в составе прецедента осуществляется поиск его вхождений в прецеденты, ранее включенные в состав кластеров. В случае обнаружения такого вхождения прецедент включается в состав кластера. При нахождении уникального прецедента, не имеющего общих сигналов ни с одним из кластеров, данный прецедент включается в состав нового кластера.

2. Определение центра кластера, в состав которого был включен прецедент, определения расстояния до центра кластера для каждого прецедента.

Центр кластера определяется значением градиента частот [8] возникновения прецедентов.

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

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

Соответственно, расстояние до центра кластера для i-го прецедента вычисляется следующим образом:

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

3. Построение неориентированного невзвешенного графа прецедентов в рамках кластера.

Вершинами графа являются прецеденты в составе кластера, связи между ними устанавливаются на основании наличия общих сигналов в составе прецедентов. Пример построения кластера представлен на рисунке 1.

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

Алгоритм C4.5 [9] предназначен для построения дерева решений с неограниченным количеством ветвей узла. Область применения алгоритма ограничена только дискретными зависимыми атрибутами, вследствие чего данный алгоритм предназначен для решения исключительно классификационных задач. Для реализации алгоритма C4.5 предъявляются следующие требования [10]:

– Каждая запись набора данных должна быть ассоциирована с одним из предопределенных классов, т.е. один из атрибутов набора данных должен являться меткой класса.

– Классы должны быть дискретными. Каждый пример должен однозначно относиться к одному из классов.

– Количество классов должно быть значительно меньше количества записей в исследуемом наборе данных.

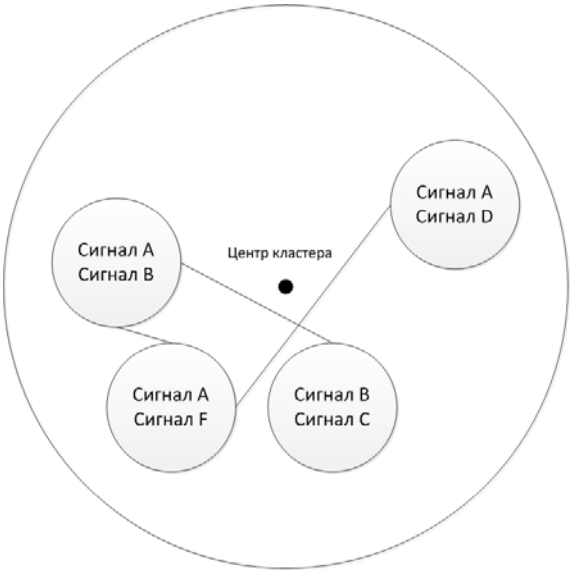

Рис. 1. – Представление кластера прецедентов в виде графа

В связи с тем, что все описанные выше требования удовлетворяются в рамках предложенного метода, алгоритм C4.5 может быть применен для построения дерева решений с целью выделения в составе кластера набора прецедентов, имеющих наибольшую взаимосвязанность.

Таким образом, предложенный метод объединения выявленных взаимосвязей между сигналами с помощью кластерного анализа позволяет распределять выявленные прецеденты на основе их взаимосвязей и частот возникновения и определять наиболее взаимосвязанные прецеденты. Данный метод может быть использован 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. Барсегян, Куприянов, Степаненко, Холод, Под ред. Барсегяна А.А. Технологии анализа данных: DataMining, VisualMining, TextMining, OLAP / 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. Клевцов С.И., Клевцова А.Б., Буринов С.В. Модель параметрической качественной иерархической оценки состояния технической системы // Инженерный вестник Дона, 2015, №3 URL: ivdon.ru/ru/magazine/archive/n3y2015/3088/.

6. Латыпова В.А. Оценка эффективности процесса обучения при наличии сложных открытых задач с помощью экспертных методов // Инженерный вестник Дона, 2016, №1 URL: ivdon.ru/ru/magazine/archive/n1y2016/3540/.

7. Гитис Л. Х. Кластерный анализ в задачах классификации, оптимизации и прогнозирования. М.: МГГУ, 2001. 103 с.

8. 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.

9. Hand, D., H. Mannila and P. Smyth, 2001. Principles of Data Mining. London: MIT Press, pp: 197-201.

10. Han, J., M. Kamber and J. Pei, 2011. Data mining: concepts and techniques. Waltham, MA, USA: Morgan Kaufmann, pp: 54-61.

References

1. Djuk V., Samojlenko A. Data mining. Uchebnyj kurs [Study Course]. SPb.: Piter, 2001. 368 p.

2. Chubukova I.A. Data mining. M.: Binom, 2008. 384 p.

3. Barsegjan, Kuprijanov, Stepanenko, Holod, Pod red. Barsegjana A.A. Tehnologii analiza dannyh [Data analysis technologies]: DataMining, VisualMining, TextMining, OLAP. 2 izd. SPb.: BHV-Peterburg, 2007. 336 p.

4. Bashmakov A.I., Dudko Ja.V. Algoritm obnaruzheniya i analiza neshtatnyh situaciy [Algorithm of emergency situations detection]. Computer science, computer engineering and management. Izhevsk: Sistemnaya inzheneriya. Nauchno-teoreticheskiy zhurnal, 2015, pp. 100-104.

5. Klevcov S.I., Klevcova A.B., Burinov S.V. Inženernyj vestnik Dona (Rus), 2015, №3 URL: ivdon.ru/ru/magazine/archive/n3y2015/3088/.

6. Latypova V.A. Inženernyj vestnik Dona (Rus), 2016, №1 URL: ivdon.ru/ru/magazine/archive/n1y2016/3540/.

7. Gitis L. H. Klasternyj analiz v zadachah klassifikacii, optimizacii i prognozirovanija [Cluster analysis in the problems of classification, optimization and forecasting]. M.: MGGU, 2001. 103 p.

8. Dubrovin B. A., Novikov S. P., Fomenko A. T. Sovremennaja geometrija metody i prilozhenija: uchebnoe posobie dlja fiziko-matematicheskih special'nostej universitetov[The methods and applications of a modern geometry : a tutorial for physical and mathematical specialties of universities]. M.: Nauka, 1986. 759 p.

9. Hand, D., H. Mannila and P. Smyth, 2001. Principles of Data Mining. London: MIT Press, pp: 197-201.

10. Han, J., M. Kamber and J. Pei, 2011. Data mining: concepts and techniques. Waltham, MA, USA: Morgan Kaufmann, pp: 54-61.