Инженерный вестник Дона, №4 (2017)
ivdon.ru/ru/magazine/archive/n4y2017/4611

Подходы к извлечению ассоциативных правил для анализа данных в нечетко-генетических системах

М.П. Кобурнеева, Е.В. Климанская

Ростовский государственный университет путей сообщения

Аннотация: В статье рассматриваются подходы к извлечению ассоциативных правил для систем гибридного искусственного интеллекта. Рассматривается известный алгоритм извлечения правил Apriori, который может применяться для обработки больших массивов количественных значений. В статье приводятся современные методы интеллектуального анализа нечетких данных: с предопределенными функциями принадлежности, алгоритмами на основе Apriori, которые обеспечивают легкий способ анализа и описания нечетких ассоциативных правил. Для работы с большими данными особенно подходящими являются алгоритмы на основе FP-деревьев. Подробно рассмотрены четыре типа нечетких генетических алгоритмов, позволяющих найти как функции принадлежности, так и нечеткие ассоциативные правила.
Ключевые слова: нечетко-генетические системы, гибридные интеллектуальные системы, ассоциативные правила, извлечение данных, интеллектуальный анализ данных

Введение

Системы гибридного искусственного интеллекта получают все большее распространение, в частности для извлечения закономерностей из больших массивов, состоящих из частично структурированных данных. Среди современного направления поисковой оптимизации наиболее современными являются генетические алгоритмы [1], искусственные нейронные сети [2] и когнитивные модели [3].

Наиболее востребованной задачей интеллектуального анализа данных является извлечение ассоциативных правил из нечетких данных Ассоциативное правило (АП) может быть представлено как “X → Y”, где X и Y это наборы элементов. Например, при анализе покупательского поведения, это правило означает что если набор элементов X покупают, то набор элементов Y покупают так же хорошо. Широко известный подход для анализа ассоциативных правил – это Apriori-алгоритм [4], который состоит из трех этапов: 1) генерация наборов-кандидатов; 2) нахождение большого

набора данных над данной минимальной поддержкой, подсчет минимальной поддержки; 3) включение ассоциативных правил над данным минимальным доверием. Поскольку традиционные ассоциативные правила учитывают отношения между элементами, их также называют бинарными ассоциативными правилами. Поскольку теория нечетких множеств обладает хорошей способностью обрабатывать количественные значения и представлять лингвистический смысл, а генетические алгоритмы хорошо подходят для оптимизации решений, то сравнительно недавно были предложены системы гибридного интеллекта, которые совмещают в себе положительные стороны обеих вышеперечисленных методов из теории искусственного интеллекта.

В работе предлагается подробное обсуждение наиболее современных методов извлечения ассоциативных правил для гибридных систем, а именно нечетко-генетических систем искусственного интеллекта.

Обзор задачи извлечения ассоциативных правил

Существующие подходы могут быть представлены в виде двух групп: с единичной минимальной поддержкой и множественной минимальной поддержкой. Под поддержкой в данном случае подразумевается некоторое пороговое граничное числовое значение. Такие методы впервые были представлены Агравалом [4] и Лью [5]. Затем несколько исследователей расширили эти подходы таксономией [6] и примерами, использующими количественные значения извлекаемые из баз данных [7]. Таким образом. данные подходы могут быть разделены на два класса:

1. алгоритмы нечеткого интеллектуального анализа данных (НИАД) с единичной минимальной поддержкой (SMSFM от single-minimum support fuzzy-mining)

2. алгоритмы нечеткого интеллектуального анализа данных с множественной минимальной поддержкой (MMSFM от multiple-minimum support fuzzy-mining).

Для методов группы SMSFM все элементы используют только одну минимальную поддержку для суждения об их важности для нечетких АП. Впрочем, при использовании этих алгоритмов, могут быть утрачены некоторые правила и элементы, значения которых проходят по верхней границе. Для преодоления этого недостатка были предложены подходы группы с MMSFM. Их главная идея заключается в том, чтобы каждый элемент имел свою величину минимальной поддержки для отражения их собственной важности.

Кроме того, поскольку элементы могут быть с таксономией, были предложены некоторые алгоритмы для разработки нечетких обобщенных АП и нечеткие многоуровневых АП.

Помимо Apriori алгоритма для разработки нечетких часто встречающихся (или популярных) элементов в виде уровней, Хан Джиавей [8] представил алгоритм frequent-pattern-tree (FP-tree, “дерево популярных предметных наборов”) и алгоритм FP-growth (адапт. русский перевод как “выращивание популярных предметных наборов”) для получения часто встречающихся наборов элементов из двоичных баз данных без генерации кандидатов. Метод состоит из двух этапов:

1. Сначала, при создании структуры FP-дерева, сохраняются только популярные единичные элементы;
2. Затем получают нечеткие популярные элементы из построенной структуры дерева FP.

Для такого типа разработки нечетких данных с концепцией дерева обработка обычно сложна из-за нечетких операторов, и, следовательно, в узлах дерева должна храниться дополнительная информация. Пападимитриу

[9] предложил алгоритм нечеткого FP-tree (FFPT), чтобы найти нечеткие ассоциативные правила, основанные на подходе pattern-growth с механизмом, подобным FP. Лин [10] также предложил несколько различных древовидных структур для создания нечетких популярных наборов предметов.

Методы извлечения ассоциативных правил в условиях неизвестной функции принадлежности нечетких данных

Алгоритмы интеллектуального анализа нечетких данных, упомянутые выше, предполагают, что функции принадлежности (ФП) уже известны заранее. Однако формы ФП могут оказывать решающее влияние на конечные результаты анализа. Разработка эффективных подходов к получению как соответствующих ФП, так и нечетких АП является задачей нетривиальной и актуальной. Одним из наиболее перспективных подходов к извлечению знаний о ФП в качестве задачи оптимизации являются генетические алгоритмы (ГА). Проблема поиска ФП и нечетких АП с ГА в иностранной научной литературе называется проблемой генетического нечеткого анализа (GFM, genetic-fuzzy mining). Способы обработки массивов данных можно разделить на два варианта:

1) обработку всех элементов вместе, комплексный подход (Integrated Genetic-Fuzzy Mining Problem for Items, IGFM);
2) обработку элементов по отдельности, подход «разделяй и властвуй» (Divide-and-Conquer Genetic-Fuzzy Mining Problem, DGFM).

Поэтому, в зависимости от типов проблем с нечеткими данными и способов обработки элементов, проблему GFM можно разделить на четыре типа, которые далее описываются достаточно подробно.

1. Комплексный подход к алгоритму нечеткого генетического анализа для элементов с единичной минимальной поддержкой (IGFM-SMS).

Подход IGFM-SMS предполагает наличие только одной минимальной поддержки для всех элементов. При этом ФП для всех элементов кодируются

в единую хромосому, а затем выводится в генетические алгоритмы. Наконец, производные функции принадлежности используются для управления нечеткими ассоциативными правилами. Было опубликовано несколько подходов к решению проблемы IGFM-SMS. Например, Хун [11] предложил генетически-нечеткий алгоритм интеллектуального анализа данных для извлечения как АП, так и ФП из количественных транзакций. Кайя [12] предложил использовать генетические алгоритмы для функций принадлежности и нечетких правил, попытавшись вывести функции принадлежности, которые могли бы достичь максимума эффективности в пределах заданного пользователем интервала минимально поддерживаемых значений. Мэтьюс [13] затем принял во внимание временную концепцию и предложил разработку временного АП с лингвистческим представлением 2-кортежей.

2. Комплексный подход к алгоритму нечеткого генетического анализа для элементов с множественными минимальными поддержками (IGFM-MMS).

В этом подходе учитывается, что различные элементы могут иметь различные свойства. Таким образом, для отражения их важности необходимы разные критерии. Например, предположим, что в наборе данных есть несколько дорогостоящих элементов, они редко покупаются из-за их высоких цен, и, таким образом, их значения поддержки являются низкими. Однако менеджер может еще быть заинтересован в этих продуктах из-за их высокой прибыли.

3. Подход "разделяй и властвуй" для элементов с единичной минимальной поддержкой (DGFM-SMS).

Преимущества IGFM в том, что они просты в использовании и с небольшими ограничением на функции пригодности GA. Однако, если количество элементов велико, для алгоритмов IGFM может потребоваться

слишком много времени, чтобы найти решение близкое к оптимальному, потому что хромосома длинная. Стратегия «разделяй и властвуй» может использоваться, когда в GFM принимается только приблизительная функция пригодности. Например, когда количество больших одноэлементных наборов используются в оценке пригодности, стратегия "разделяй и властвуй" становится отличным выбором для решения этой проблемы, поскольку каждый элемент обрабатывается индивидуально.

4. Подход "разделяй и властвуй" для элементов с множественной минимальной поддержкой (DGFM-MMS).

В данном случае, проблема может рассматриваться как комбинация IGFM-MMS и DGFM-SMS. Таким образом, структура проблемы DGFM-MMS может быть легко разработана из предыдущих рамок для IGFM-MMS и DGFM-SMS. Поскольку DGFM-MMS является сложным алгоритмом, есть весьма ограниченное количество источников по теме нахождения минимальных поддержек и ФП элементов для нечетких АП. Так как в практических задачах количество АП велико и не может быть легко закодировано в хромосоме, большинство предложенных методов сначала изучали ФП для элементов, а затем, в соответствии с полученными ФП выводили нечеткие АП.

В соответствии с типами правил подходы так же можно разделить на четыре типа, включая нечеткие АП, нечеткие обобщенные АП, нечеткое взвешенное АП и нечеткое временное АП. В предыдущих подходах больше внимания уделяется нечетким АП и взвешенным нечетким АП.

На первом этапе генетический процесс используется для получения функций принадлежности для нечетких значений. На втором этапе нечеткие ассоциативные правила запускаются методом нечеткого поиска на основе производных функций принадлежности.

Заключение

В данной статье был проведен сравнительный анализ подходов к извлечению знаний, в данном случае ассоциативных правил из массивов частично структурированных данных. Среди них были методы, основанные на Apriori алгоритме, на алгоритме FP-деревьев и генетическом алгоритме. Главная идея этих подходов сосредоточена на применении теории нечетких множеств для обработки количественных значений. В подходах на основе Apriori количественные значения преобразуются в нечеткие множества в соответствии с предопределенными функциями принадлежности. Затем создаются нечеткие популярные наборы элементов и нечеткие АП на основе процесса выполнения алгоритма Apriori. Время выполнения такого анализа довольно продолжительно. Более быстрый вариант основан на алгоритме FP-tree. В целом функции принадлежности могут предоставляться экспертами, однако это не всегда применимо. Для автоматического получения соответствующих функций принадлежности предлагается использование нечетких генетических алгоритмов. Определены четыре типа подходов к нечеткому генетическому анализу, включая IGFM-SMS, IGFM-MMS, DGFM-SMS и DGFM-MMS.

Работа выполнена при финансовой поддержке РФФИ, проект № 16-01-00597-а.

Литература

1. Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы. М.: ФИЗМАТЛИТ, 2010. 368 с.

2. Пучков Е.В. Сравнительный анализ алгоритмов обучения искусственной нейронной сети // Инженерный вестник Дона, 2013, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2135

3. Гинис Л.А. Развитие инструментария когнитивного моделирования для исследования сложных систем // Инженерный вестник Дона, 2013, № 3, URL: ivdon.ru/ru/magazine/archive/n3y2013/1806

4. Agrawal, R., Imielinski, T., Swami, A. Database mining: A performance perspective. IEEE Trans. Knowl. Data Eng. 5, 1993. pp. 914–925

5. Liu, B., Hsu, W., Ma, Y. Mining association rules with multiple minimum supports. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,1999. pp. 337–341

6. Tseng, M.C., Li, W.Y. Efficient mining of generalized association rules with non-uniform minimum support. Data Knowl. Eng. 62(1), 2007. pp. 41–64

7. Lee, Y.C., Hong, T.P., Wang, T.C. Multi-level fuzzy mining with multiple minimum supports. Expert Syst. Appl. 34(1), 2008, pp. 459–468

8. Han, J., Pei, J., Yin, Y., Mao, R. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min. Knowl. Disc. 8, 2004, pp.53–87

9. Papadimitriou, S., Mavroudi, S.: The frequent fuzzy pattern tree. The WSEAS International Conference on Computers, 2005, pp.145-166

10. Lin, C.W., Hong, T.P., Lu, W.H. An efficient tree-based fuzzy data mining approach. Int. J. Fuzzy Syst. 12, 2010. pp. 150–157

11. Hong, T.P., Chen, C.H., Wu, Y.L., Lee, Y.C. A GA-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions. Soft Comput. 10 (11), 2006. pp. 1091–1101

12. Kaya, M. Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules. Soft Comput. 10(7), 2006, pp. 578–586

13. Matthews, S.G., Gongora, M.A., Hopgood, A.A., Ahmadi, S. Temporal fuzzy association rule mining with 2-tuple linguistic representation. The IEEE International Conference on Fuzzy Systems, 2012. pp. 1–8

References

1. Gladkov L.A., Kurejchik V.V., Kurejchik V.M. Geneticheskie algoritmy. [Genetic Algorithms] M.: FIZMATLIT, 2010. p.368

2. Puchkov E.V. Inženernyj vestnik Dona (Rus), 2013, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2135

3. Ginis L.A. Inženernyj vestnik Dona (Rus), 2013, №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1806

4. Agrawal, R., Imielinski, T., Swami, A. IEEE Trans. Knowl. Data Eng. 5, 1993. pp. 914–925

5. Liu, B., Hsu, W., Ma, Y. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999. pp. 337–341

6. Tseng, M.C., Li, W.Y. Data Knowl. Eng. 62(1), 2007, pp. 41–64

7. Lee, Y.C., Hong, T.P., Wang, T.C. Expert Syst. Appl. 34(1), 2008, pp. 459–468

8. Han, J., Pei, J., Yin, Y., Mao, R. Data Min. Knowl. Disc. 8, 2004, pp.53–87

9. Papadimitriou, S., Mavroudi, S. The WSEAS International Conference on Computers, 2005, pp. 145-166.

10. Lin, C.W., Hong, T.P., Lu, W.H. Int. J. Fuzzy Syst. 12, 2010, pp. 150–157

11. Hong, T.P., Chen, C.H., Wu, Y.L., Lee, Y.C. A GA-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions. Soft Comput. 10 (11), 2006, pp. 1091–1101.

12. Kaya, M. Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules. Soft Comput. 10(7), 2006, pp. 578–586.

13. Matthews, S.G., Gongora, M.A., Hopgood, A.A., Ahmadi, S. The IEEE International Conference on Fuzzy Systems, 2012, pp. 1–8.