

Применение методов Data Mining в социологии.

Хавенсон Т.Е.; ГУ-ВШЭ, Москва

Методы Data Mining¹ завоевали огромную популярность в биологии, медицине, в интернет-технологиях, в принятии решений в бизнесе. Но пока что еще не получили должного распространения в социологии, особенно в российской.

Одно из наиболее популярных определений Data Mining дано в 1996г. одним из основателей этого направления Г. Пятецким-Шапиро: Data Mining – это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных для интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. [1, 58].

Data Mining оформилось как направление в анализе данных во второй половине XX века. Причем появление этого направления необязательно связано с разработкой именно новых методов, многие методы были известны и раньше. Скорее DM – это зарождение новой философии, нового взгляда на анализ данных. Одним из первых, кто заговорил об этом, был известный ученый в области математической статистики – Джон Тьюки (John Tukey). В 1962 г. он написал статью, которая называлась "Будущее анализа данных" (The future of data analysis), в которой изложил основные идеи новой тенденции. Тьюки говорил о том, что точность и строгость математических основ статистики не помогают в решении реальных жизненных проблем и что надо дать данным говорить самим за себя. Отличие DM от классических методов математической статистики, используемых в анализе данных, в том, что в DM во главу угла ставятся сами данные. Получение новых знаний производится без предъявления требований со стороны метода к данным. [2, 5; 3]

Как было сказано выше, методы Data Mining приобретают в последнее время все большую популярность. Для этого есть несколько причин.

Во-первых, сейчас особо остро стоит задача обработки именно больших объемов информации. В последние десятилетия накопление информации в самых разных сферах общественной жизни стало обычной практикой, в результате появилось огромное количество многомерных баз данных, содержащих тысячи записей. И здесь методы Data Mining имеют преимущество, потому что они автоматизируют процесс от начала до конца, то есть включают в себя и подготовку данных к анализу, и собственно

¹ Чаще всего Data Mining переводится на русский язык как "добыча данных", но также в русскоязычной литературе уже сложилась традиция использования англоязычного термина. Мы будем придерживаться этой традиции, а также пользоваться сокращением DM.

реализацию конкретных алгоритмов, и выдачу результатов в удобном пользователю виде.

В качестве второй причины популярности Data Mining можно назвать то, что среди всего арсенала методов DM практически все методы подходят для анализа номинальных данных. Сейчас уже все большее количество ученых, аналитиков и других специалистов в области изучения социальной реальности сходится в том, что мир дискретен, а социальный мир в особенности. Само существование переменной как некоего непрерывного континуума, присущего всем объектам, ставится под сомнение. Социальные явления, события или другие объекты лучше описываются с помощью набора каких-либо характеристик, а их сочетания могут определять типы респондентов.

Основные задачи, решаемые методами Data Mining:

Классификация – это отнесение объектов к одному из заранее известных классов. Кластеризация – это группировка объектов на основе характеристик, описывающих сущность этих объектов. Регрессия, в том числе задачи прогнозирования, – устанавливает зависимость выходных переменных от входных; позволяет определить по известным характеристикам объекта значение некоторого его параметра. Поиск ассоциативных правил – выявление закономерностей в виде правил "Если А, то В" между связанными событиями. [1,59-60; 4].

Основные методы Data Mining:

В целом методы DM по-прежнему во многом основываются на классических принципах разведочного анализа данных и построения моделей. Среди них – нейронные сети, деревья решений, методы ограниченного перебора, генетические алгоритмы, кластерные модели, комбинированные методы и другие. [5, 6]

СПИСОК ЛИТЕРАТУРЫ:

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining. Visual Mining. Text Mining. OLAP. СПб.: БХВ-Петербург, 2008
2. Miller, T. W. Data and text mining. Prentice Hall, 2005
3. Handbook of Computational Statistics. Concepts and Methods. Ed. J.E. Gentle, W. Härdle, Y. Mori. Ch. 13. Data and Knowledge Mining
4. Data Mining – добыча данных // <http://www.basegroup.ru/tasks/datamining.htm>
5. Электронный учебник по статистике. Москва, StatSoft Inc., 2001. <http://www.statsoft.ru/home/textbook/default.htm>.
6. Чубукова И.А. Data Mining <http://www.intuit.ru/department/database/datamining/>