А.А.Давыдов

руководитель группы «Анализ социальных систем»
Института социологии РАН,
Вице-президент Российского общества социологов

KDD Cup 2009: полезный опыт победителя

Ключевые слова: системная социология, анализ данных, Knowledge Discovery and Data Mining, KDD Cup 2009.

Введение

Традиционно, в рамках ежегодных международных ACM SIGKDD Conference on Knowledge Discovery and Data Mining [1] проводится KDD Cup [2] - соревнование аналитиков данных с использованием компьютерных систем Knowledge Discovery and Data Mining (KDD) [3], предназначенных для индустриального анализа и моделирования данных. На KDD Cup ставится массовая (часто встречающаяся) в практике исследовательская задача в области анализа данных и прогнозирования, которую решают несколько тысяч исследователей из многих стран мира. Так, например, в KDD Cup 2009 [4] участвовало 7865 участников, представляющих 1299 исследовательских команд, из 46 стран мира. Особенностью KDD Cup являются следующие обстоятельства: участвует много студентов и молодых ученых, не используются суперкомпьютеры и High-performance Knowledge Discovery and Data Mining systems (высокопроизводительные KDD), в частности, Distributed Knowledge Discovery and Data Mining (распределенные в географическом пространстве KDD), основанные на технологии Knowledge Grid and Grid Intelligence (KGGI) [3], ограниченно используются коммерческие компьютерные системы Knowledge Discovery and Data Mining (KDD).

В данном сообщении кратко приводятся методология и методика победителя KDD Cup 2009 [4], которые отражают успехи студентов и молодых ученых в области анализа и моделирования данных. Представленные в данном сообщении результаты являются полезными для российских студентов-социологов и молодых ученых, занимающихся анализом эмпирических данных, в частности,

больших массивов гетерогенной (количественные и качественные переменные) «зашумленной» (пропуски в данных и погрешности измерения) социологической информации, например в Базах данных (БД) результатов опросов общественного мнения, в Интернете, корпоративных Базах данных потребителей (клиентов) и т.д.

KDD Cup 2009: конкурсная задача «Fast Scoring on a Large Database»

Конкурсная задача относилась к классу массовых (распространенных) содержательных задач Customer Relationship Management (CRM), которая является ключевым элементом современных маркетинговых стратегий. Конкурсная задача состояла в том, чтобы как можно быстрее и точнее предсказать поведение потребителей для трех стандартных в маркетинге переменных, а именно Churn (to switch provider), Appetency (to buy new products or services), Up-selling (to buy upgrades or new options proposed to them).

Тестовый эмпирический материал включал в себя 100 000 случаев (объектов). 50 000 случаев – выборка для построения моделей и 50 000 случаев для тестирования моделей. Количество переменных - 15 000 переменных (категориальных и количественных, с пропусками, «зашумленных» и т.д.) из большой Базы данных French Telecom company Orange - одного из мировых лидеров в области телекоммуникационных услуг (более 170 млн. потребителей). В этой связи напомним, что быстрый анализ и прогнозирование в больших Базах данных (БД) - стандартная задача анализа, прогнозирования и компьютерного моделирования в информационном обществе, основанном на знаниях [5-6]. Таким образом, конкурсная задача состояла в том, чтобы быстро (не более, чем за пять дней) найти эмпирические правила на массиве из 100000 случаев по 15 000 переменным, для точного предсказания значений трех переменных. Это стандартная исследовательская задача Predictive Analytics, которую некоторые команды решили за несколько часов.

KDD Cup 2009: методология и методы победителя

Победителем KDD Cup 2009 стала команда из 11-ти аналитиков данных из IBM T.J. Watson Research Center (http://www.watson.ibm.com/index.shtml) - одного из мировых лидеров в области разработки инновационных методов анализа информации и программного обеспечения. Члены команды IBM имели четко обозначенные функциональные обязанности, а именно, анализ данных,

программирование, IT-технологии, заполнение пропусков в данных и т.д. и применяли следующие компьютерные решения, методологию и методики прогнозирования.

Использовался Linux кластер, состоящий из 9 задействованных параллельно персональных компьютеров, каждый из которых имел по два процессора и 3 Гб оперативной памяти.

Использовались следующие языки программирования: C/C++, Java, Matlab. В этой связи отметим, что данные языки программирования входят в число учебных дисциплин, которые изучают социологи за рубежом [7].

Использовалась методология параллельного и распределенного анализа данных, моделирования и прогнозирования. Также использовалась методология и методика Machine Learning (машинного обучения), генерации и селекции моделей Ensemble Selection - «ансамбля» (множества) моделей. Так, например, в данном исследовании генерировались «ансамбли» из 700-1200 моделей и использовалась десятикратная перекрестная валидизация моделей.

Основные пакеты и библиотеки методов анализа - FEST package, BBR package, SNoW package, SVMPerf, LibLinear, LibSVM, Weka, KNN и т.д. Автор рекомендует российским социологам ознакомиться с интересным и полезным пакетом Weka (http://www.cs.waikato.ac.nz/ml/weka/), который предназначен для Ensemble Selection. Отметим, что перечисленные пакеты работали параллельно и распределенно в единой вычислительной системе. В данном исследовании системно использовались статистические методы, эвристические алгоритмы и т.д., что соответствует принципам системного анализа данных [8], которые используются в компьютерных системах Knowledge Discovery and Data Mining (KDD) [3].

Preprocessing and feature construction:

- Normalizations (for numerical variables)
- Replacement of the missing values
- Discretization (for numerical variables)
- Principal Component Analysis

Classification:

- Base classifier
- Decision tree, stub, or Random Forest
- Linear classifier (Fisher's discriminant, SVM, linear regression)
- Non-linear kernel method (SVM, kernel ridge regression, kernel logistic regression)
- Naive Bayes
- Bayesian Network (other than Naive Bayes)
- Nearest neighbors

Loss function:

- Hinge loss (like in SVM)
- Square loss (like in ridge regression)
- Logistic loss or cross-entropy (like in logistic regression)
- Exponential loss (like in boosting)

Regularizer:

- One-norm (sum of weight magnitudes, like in Lasso)
- Two-norm (||w||^2, like in ridge regression and regular SVM)

Ensemble method:

- Boosting
- Bagging
- Other ensemble method

Model selection/hyperparameter selection:

- The on-line feed-back on 10% of the test set was used
- K-fold or leave-one-out cross-validation (using training data)

В результате использования данной методологии и методики, команде победителей KDD Cup 2009 удалось, менее чем за два дня, предсказать поведение потребителей с вероятностью правильного предсказания 85%.

Заключение

Для российских студентов – социологов и молодых ученых, занимающихся анализом данных, в 2010 году представится возможность проявить себя и получить денежный приз на следующем KDD Cup 2010 [9], который проводится в рамках the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-2010). Washington, DC, USA, 25-28 July 2010 [10]. Победа в KDD Cup дает широкие возможности для международного признания и последующей успешной научной карьеры. Так, например, руководитель команды - победителя KDD Cup 2009 Alexandru Niculescu-Mizil, стал руководителем KDD Cup 2010 [9].

СПИСОК ЛИТЕРАТУРЫ

1. http://www.sigkdd.org/conferences.php
2. http://www.sigkdd.org/kddcup/index.php
3. Давыдов А.А. Knowledge Discovery and Data Mining в системной социологии. М.: ИС РАН, 2009. (http://www.isras.ru/Davydov_Knowledge.html)
4. http://www.kddcup-orange.com/prizes.php
5. Давыдов А.А. Системная социология: Ultra-Large-Scale Holistic Simulation. М.: ИС РАН, 2009. (http://www.isras.ru/index.php?page_id=1008)
6. Давыдов А.А. Системная социология: Data Warehousing. М.: ИС РАН, 2009. (http://www.isras.ru/index.php?page_id=1012)
7. Давыдов А.А. Системная социология: языки программирования. М.: ИС РАН, 2009. (http://www.isras.ru/index.php?page_id=1075)
8. Давыдов А.А. Системная социология: введение в анализ динамики социума. М.: ЛКИ, 2007.
9. http://www.kdd.org/kdd2010/kddcup.shtml
10. http://www.kdd.org/kdd2010/index.shtml