Capital Crunch: Predicting Investments in Tech Companies

Zifei Shan, Haowen Cao and Qianying Lin
Department of Computer Science, Stanford University
{zifei,caohw,qlin1}@stanford.edu

Keywords

Machine Learning, Data Mining, Link Prediction, Startups

ABSTRACT

For many start-ups, lack of investment and capital has become the bottleneck for development. This phenomenon inspires us to use machine learning algorithms to find patterns in investment behavior from major investors. We plan to use various domain-specific features to predict which investors would potentially invest in a particular company . This would not only reveal important information about investment strategies and behaviors of investors, but also give startups ideas on where to seek potential investment and how to adjust their strategies so as to attract potential investors.

Our work is grounded in CrunchBase, an accessible knowledge base that maintains full records of company and people information.

There are two primary goals of our work:

- To predict whether an investor would invest in a particular start-up based on textual, topological and domain-specific signals from both the investor and start-up. - To analyze and reveal the factors that would prompt an investor to invest in startups so as to shed light on the adjustments the start-ups could make to attract more investments.

<!– Our goal is to infer investment events in technology companies using various domain-specific features. This work can potentially cast insights to understanding factors affecting investment behaviors, making strategies for companies and investors, and mining interesting patterns happening in the market. Our work is grounded in CrunchBase, an accessible knowledge base that maintains full records of company and people information.

We ask questions such as: what factors play the most critical role in investments? Can we infer whether an investor will invest a certain company by textual, topological and domain-specific signals? –>

1. DATASET

We use data from *CrunchBase.com*, one of the biggest databases about information of companies. The current CrunchBase dataset includes 214,290 companies and 286,659 people.

1.1 Accessing data

CrunchBase provides indexing data and an API for full access of their data, yet the API has limited throughput. Due to the limitation, by now we would like to sub-sample the dataset, and we may get the full dataset in the future.

For data sub-sampling, a possible strategy is random sampling, where we randomly take out a certain portion of the data. However, it would have the potential drawback that we might not achieve consistency amongst different parties. For example, if we sampled Facebook but not some other companies where Mark Zuckerberg has been CEO, then we would lose some information with regard to the network. Therefore we propose to adopt the strategy where we start with a “seed set” of companies, and get all related people and organizations in an iterative way to grow the network.

1.2 Data format

In terms of data format, CrunchBase provides a complete index of the people and organizations, which includes a unique identifier that we can use to get detailed data for people and organizations via API. The detailed data format is demonstrated below.

The people data are like this:

```
"data": {
  "uuid": "a01b8d46d31133337c34aa3ae9c03f22",
  "properties": {
    "bio": ...
    "last_name": ...
    "first_name": ...
    ...
  }
  "relationship": {
    "degrees": {...}
    "experiences": {...}
    "news": {...}
    ...
  }
}
```

The organization data for startups are like this:

```
"data": {
  "properties": {
    "description": {...}
    "founded_on": {...}
    "name": {...}
    "number_of_employees": {...}
  }
  "relationships": {...}
  "borad_members_and_advisors": {...}
  "acquisitions": {...}
  "competitors": {...}
  ...
}
```

With these data as input, we construct models and run machine learning algorithms to get predictions on investments. The section below articulates our proposed model.

2. PROPOSED MODEL

2.1 Data Model

The crunchbase dataset has a variety of entities: **organization**, **person**, **product**, etc. There are also different relations including **investment**, **acquisition**, **degree**, **founder**, etc.

For simplification, we categorize **organizations** into **startups** and **investors**, and we care about predicting **investment** relationship between them.

The data model is defined below:

- $Startup(startupId, [attributes...])$
- $Investor(investorId, [attributes...])$
- $Investment(investorId, startupId, isTrue)$

Where we use features in $Startup$ and $Investor$ entities to predict $Investment$ relations.

2.2 Problem definition

Given the full $Startup$ relation and $Investor$ relation, predict $isTrue$ value in $Investment$ table, which determines if any given investor invests a startup.

TODO

2.3 Labeled data

We take ground truth investments in CrunchBase as positive training examples

TODO

2.4 Proposed Features

A rich set of features can be applied to predict investments. They may include:

- textual features: company descriptions, biography of people.
- TODO

2.5 Baseline and Oracle

A naive baseline model would be a random predictor that predict random

We propose to apply a factor graph model that correlates features across this graph. Specifically, TODO

TODO baseline: logistic regression

3. EVALUATION

We want to use the information from people and organizations, finding the connectionsïijŇ common points and all the relationship we can get from the data to learn a graph. Using this graph, we want to predict the probability of the investors a company, and what kind of investors, which exactly investor will invest on a certain kind of company.The data we get from crunchbase will be separated into two parts: training and testing, to help us evaluate the behavior of our model and help us chose good predictors and relationship between nodes (including people and organizations). Therefore we could evaluate if our model is trained well and use the information we get properly.

4. CHALLENGES

Data sparsity. In our dataset, some investors might be totally unrelated to some companies. A naive predictor would just predict them to be "not investing", but we would like to delve deeper to see the possible subtleties.

The need to integrate the various parties and their relationships into one problem. For example we have start-ups, venture capital companies, and people (e.g. the founder of a start-up), we need to use a model so as to accurately capture their relationship.

4.1 Topics to Address Challenges

Natural Language Processing : We would use the Stanford NLP to process the short descriptions so as to obtain information such as which area the start-up is focusing on.

Probabilistic Graphical Models: PGM would be adopted to examine the relationship between founders, companies, and investors.

TODO

Factor Networks: Factor networks would be used to model the whole dataset. For example, degree would be an edge to connect a founder of the start-up and a founder of a venture capital company.

5. RELATED WORK

In the paper Recommending Investors for Crowdfunding Projects [1], the author discussed a methodology to match proposals from start-ups to the potential investors on Kickstarter with linear regression, SVM-linear, SVM-poly and SVM-RBF, with an accuracy rate of 82% for static data features and 73% for dynamic data features. Thought their features are mostly updates made to the tweets , number of comments and so on, we could expand the feature set to other features such as the education background of the investor, the area of investment the investor usually specializes in, etc.

Another paper, Predicting new venture survival: an analysis of "anatomy of a start-up." [2], Gartner, Starr and Bhat, used a discriminant analysis to classify the potentially successful company and unsuccessful companies. Their feature sets are worth noting, including individual characteristics of the entrepreneurs, the efforts by entrepreneurs (i.e. whether they actively look for resources and help), degree of innovation and so on. Though this paper is more on the social science side, we would like to scrutinize the feature sets so as to explore more meaningful and insightful features. For example, we could extend individual characteristics to how many start-ups the CEO has founded and their past histories. In this way, we would be able to obtain a richer set of data.

6. REFERENCES

[1] J. An, D. Quercia, and J. Crowcroft. Recommending investors for crowdfunding projects. In *Proceedings of the 23rd international conference on World wide web*, pages 261–270. International World Wide Web Conferences Steering Committee, 2014.

[2] W. Gartner, J. Starr, and S. Bhat. Predicting new venture survival: an analysis of âĂIJanatomy of a start-up.âĂİ cases from inc. magazine. *Journal of Business Venturing*, 14(2):215–232, 1999.