Project Report:

Who is more likely to gain a large number of citations?

5130309465 孙元璞

1. Background&Meaning

As the volume of literature collection rises at a tremendous speed, it’s necessary to find a method that can find those original literature that are more likely to become popular. In this project, I use the count of citations as a measurement for the popularity among researchers and try to use some simple features to predict the count of citations that a paper will receive. By comparing these counts, I can find those papers that will be popular among researchers but have not be frequently cited yet.

2. Idea&Method

First I try to find those influence factors which are important as well as easy to find. For example whether an article is original or not is important, but it cost lots of time to read the article and judge its original or not, so this kind of features will be excluded at the very beginning.

2.1 Problem Definition

Citations. Given the literature corpus D, the citation counts (C_T (.)) of a literature article $d \in D$ is defined as:

$$\text{citing}(d) = \{d' \in D : d' \text{ cites } d\}$$

$$C_T(d) = |\text{citing}(d)|$$

Learning task: Given a set of article features, $\vec{x} = x_1, x_2, \ldots, x_n$, my goal is to learn a predictive function f to predict the citation counts of an article d after a give time period Δt(in this project Δt is set as 5 years):

$$f(d|\vec{x}, \Delta t) \rightarrow C_T(d|\Delta t)$$

2.2 Feature Definition

2.2.1 H-Index

H-index h means a researcher has at most h papers that are cited at list h times. The h-index is useful which attempts to measure both the productivity and impact of the published work of a scientist. By the definition,it can be considered as the reference of the possibility of whether a paper is important in all the papers which belongs to the same author.The index is based on the set of the scientist’s most cited papers and the number of citations received in others’ publications. Besides,h-index has been proved to have predictive power of scientific output and impact of a researcher[1]. Therefore, I choose h-index as a candidate feature to predict citation counts.

2.2.2 Paper Count

Sometimes larger paper count means the author is more active so that more popular among researchers.I chose their paper counts in the recent 5 years.As those who haven’t appear 5 years ago,I consider that it also means he is more likely to have no fame.So this is also equitable to some degree.

2.2.3 Venue Rank

Like authors, venues also have academic reputations.Here I choose the CCF conference and journal ranking to distinguish those conferences and journals as class A,B,C .For new venues, I use the minimum feature values instead of N/A: anything has a start.

2.3 Predictive Model

I choose the Support Vector Regression to solve this problem.Statistical Learning Theory has provided a very effective framework for classification and regression tasks involving features. Support Vector Machines (SVM) are directly derived from this framework and they work by solving a constrained quadratic problem where the convex objective function for minimization is given by the combination of a loss function with a regularization term (the norm of the weights). There are two main categories for support vector machines: support vector classification (SVC) and support vector regression (SVR). SVM is a learning system using a high dimensional feature space. It yields prediction functions that are expanded on a subset of support vectors. The model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close to the model prediction. Support Vector Regression is the most common application form of SVMs. An overview of the basic ideas underlying support vector machines for regression and function estimation has been given in details in [2].

3. EXPERIMENTS AND EVALUATION

IN this project,I randomly choose 1000 papers from the MAG (Microsoft Academic Graph)to complete the SVM model and use it to predict 8 papers from 8 different researchers to test and verify whether this method is reliable or not.The following table show some results about those 8 researchers(from 2012 to 2016).

Name	Tom Mitchell	Yong Yu	Jiawei Han	Kai Li	Yuanyuan Zhou	Dina Katabi	Garth Gibson	Michael I Jordan
Paper count	73	66	283	131	39	79	51	186
H-index	76	53	159	80	56	64	63	146
Citation Count (2017)	2941	1801	13025	4124	1071	2347	1170	14146
Predict Count	3466	1635	14353	4527	779	2420	937	16323

The result contains a large deviation.Obviously,to get a more accurate model,more detail features should be added and more samples should be used to train the predict model.

As expected,H-index is the most important independent variable in these features,so if you have no time to judge which author’s paper may help you,H-index can help you a lot.

References

[1]J. Hirsch. Does the h index have predictive power?Proceedings of the National Academy of Sciences,104(49):19193, 2007.

[2]A. Smola and B. Schölkopf. A tutorial on support vector regression. Statistics and computing, 14(3):199–222, 2004.