Classifying Search Advertisers

By Lars Hirsch   (Sunet ID : lrhirsch)

Summary

Multinomial Event Model and Softmax Regression were applied to classify search marketing advertisers into industry verticals using advertiser spend data across keywords and keyword categories. Using Multinomial Event Model, an error rate of 34% was achieved. Using Softmax Regression, an error rate of 38% was achieved.

Background

Advertising works by making the customer aware of the product and by focusing on customer’s need to buy the product. Globally, advertising has become an essential part of the corporate world. Therefore, companies allot a huge part of their revenues to the advertising budget. With the overwhelming penetration of the internet, online advertising consumes a major part of this budget. In United States alone Internet advertising revenues amounts to $40 billion.

Like other advertising media, online advertising frequently involves both a publisher, who integrates advertisements into its online content, and an advertiser, who provides the advertisements to be displayed on the publisher’s content. Most of Search Advertising is built around the pay per click business model, where advertisers Keywords are matched to user Queries, and advertisers only pay for ads that are clicked by a user. With this business model it becomes crucial that the correct Keywords are selected for each Query, and relevant ads are displayed onto a user's screen. An integral part of deciding which ads are relevant, is to understand the advertisers. Hence any system that would help identify advertisers would provide a valuable edge in the endeavor to target relevant advertisement to users.

Currently at Microsoft Bing Ads, our advertiser base differ considerably in terms of their business. The approach that has been adopted is to hire vendors that would manually classify our large advertisers with no classification available for small advertisers. However this approach over time has proved expensive and is not scalable as our advertiser base is growing. The unavailability of classification for smaller advertisers is limiting the effectiveness of our marketing efforts and is a blind spot with regards to knowledge of our customer base. This becomes very crucial as we move into developing markets.

Two supervised learning models were developed to classify advertisers into different industry verticals (labels). The training sets used consisted of the manually classified advertisers that we currently possess and use features such as spend weighted distribution of keywords and query categories - note that query (supply) taxonomy is different from the advertiser (demand) taxonomy. A single query category could be relevant for multiple advertiser categories (for example, "soccer" which has query class "sports", could be relevant for advertiser categories such as "entertainment", "sporting goods" etc.). Also, while query intent and therefore category is often ambiguous (for example "jaguar"), advertiser category is not.

Execution

Classifying advertisers is similar to classifying email into spam/non-spam in the sense that we are looking for particular “words” (aka Keywords in a search advertising context) that identify the subject as belonging to a particular category. Here, we are looking at identifying the correct Industry Vertical, e.g. Automotive, Retail, Financial Services etc., which an advertiser belongs to.

Two types of features were used:

- The top 10, 50, 100, 1000 and 10,000 Keywords advertiser bid for, by overall spend for each advertiser vertical, were selected. Then, percentage of spend per advertiser across each of these keywords were calculated and used as features.
- Keywords are categorized based upon the meaning of the word according to an internal dictionary known as CHE. This is a two-level categorization, with 37 top levels and 1237 branch levels. Percentage of advertiser spend across CHE level 1 and CHE level 2 were calculated and used as features as well.

Keywords and CHE categories were used independently and combined:

- *n* top Keywords (i.e. *n* features)
- CHE level 1 (37 features)
- CHE level 2 (1393 features)
- *n* top Keywords with CHE level 1
- *n* top Keywords with CHE level 2
- CHE level 1 with CHE level 2
- *n* top Keywords with CHE level 1 and CHE level 2

Data was collected for all currently manually classified advertisers in the system with a spend >$0 for the period extracted; initially 15 day periods, then 7 (due to the time required to extract the data for 15 days) and 30 day periods were tried as well to check if the model was sensitive to the data collection period (it was for certain feature sets). Data was extracted from the log files using Microsoft’s internal MapReduce cluster known as Cosmos/Scope.

Since this is a multi-class classification problem with a high-dimensionality feature vector, and with the similarity with spam classification in mind, Multinomial Event Model – a variation of Naïve Bayes Classifier - seemed to be a natural initial choice of algorithm. Since Naïve Bayes rely upon the unrealistic assumption of independence of features, an alternative classification algorithm – Softmax Regression – was tried as well.

Initially, the algorithms were run in Octave, but with bigger and bigger datasets, up to several hundred MB of data within some training set, it was eventually not feasible to run the algorithms in Octave. Therefore, a custom Multinomial Event Model was implemented using MapReduce, and run using Cosmos/Scope. Softmax Regression was run on the smaller datasets using R (it was not run on the larger datasets).

For all models, a 20% simple cross-validation sample was use to test each model.

Multinomial Event Model

Since the features represent the percentage of spend for a given keyword or CHE vertical, they will usually be << 1, and the convention of using a Laplace smoothing factor of 1 in the numerator will not be ideal. Multinomial Event Model was applied, using a Laplace smoothing factor equal to the lowest value of any feature in the feature set (and number of features times this lowest value in the denominator). The ϕ vector and prior probability was calculated for each possible class (advertiser vertical) using the training set, and the log likelihood was calculated on the cross-validation set.

In other words,

$$\phi_{k|y=c} = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = c\}\, x_k^{(i)} + \min(x)}{\sum_{i=1}^{m}\sum_{j=1}^{n} 1\{y^{(i)} = c\}\ x_j^{(i)} + \min(x) * |V|}$$

Where k is a feature, c is a class (i.e. advertiser vertical), $x_k^{(i)}$ is the percentage of spend by advertiser i on feature k (which can be a keyword or a keyword category i.e. CHE). n is number of features, so $\sum_{j=1}^{n} x_j^{(i)} = 1$ for any i if an exhaustive and non-overlapping list of features are included (which is the case for CHE but not for KWs or any combination of feature types such as KWs + CHE), since in that case spend will add up to 100%.

Finally, prediction was run on the cross-validation sample, and the advertiser was assigned to the class (vertical) with the highest log likelihood value given the feature set of the training example, the prior and the ϕ vector for the class.

The best performing Multinomial Event Model combined spend distribution (for each advertiser) across CHE1 with spend distribution across CHE2 (table 1), resulting in an error rate of 34% on the cross-validation set. See table 3 for error rates for other feature combinations.

Multinomial Event Model CHE1+CHE2

Error rate 34%

					Predicted					*Total*	
		Autos	Edu	Finance	Health	Other	Retail	Tech	Travel	*actual*	Recall
Actual	Autos	**38**				10	8	1		*57*	67%
	Education	2	**34**	2	3	37	2	2		*82*	41%
	Finance	2		**94**	1	24	10	2	1	*134*	70%
	Health		1		**45**	26	12			*84*	54%
	Other	2	6	5	13	**205**	58	12	4	*305*	67%
	Retail	1	1	6	7	76	**320**	3	4	*418*	77%
	Technology				1	35	9	**33**		*78*	42%
	Travel	1				26	8	2	**47**	*84*	56%
	Total predicted	*46*	*42*	*107*	*70*	*439*	*427*	*55*	*56*		
	Precision	83%	81%	88%	64%	47%	75%	60%	84%		

***Table 1**: Cross-validation of Multinomial Event Model with CHE1 and CHE2 as features.*

Softmax Regression

Softmax Regression using stepwise feature selection with AIC was run on the CHE level 1 dataset, producing a model with an error rate of 38% (table 2).

Softmax Regression CHE1

Error rate 38%

		Predicted									
		Autos	Edu	Finance	Health	Other	Retail	Tech	Travel	*Total actual*	Recall
Actual	Autos	**41**	2	0	0	7	5	0	0	*55*	75%
	Education	0	**28**	1	4	21	3	3	0	*60*	47%
	Finance	1	3	**77**	2	20	7	4	2	*116*	66%
	Health	0	0	1	**49**	19	11	0	0	*80*	61%
	Other	6	10	10	22	**205**	92	15	10	*370*	55%
	Retail	4	3	3	13	83	**279**	9	2	*396*	70%
	Technology	0	3	1	3	31	14	**33**	0	*85*	39%
	Travel	1	0	0	1	11	5	1	**49**	*68*	72%
	Total predicted	*53*	*49*	*93*	*94*	*397*	*416*	*65*	*63*		
	Precision	77%	57%	83%	52%	52%	67%	51%	78%		

***Table 2**: Cross-validation of Softmax Regression using advertiser spend distribution across CHE1 as features.*

Using PCA, the CHE level 2 dataset was reduced from 1393 features to 75 principal components, explaining 50% of the variance in the original dataset. This was necessary to run Softmax Regression. Running prediction on the training set, the error rate was 0%, while on the cross validation set it was 43% - clearly an issue of overfitting. Again, stepwise feature selection with AIC was run, but the algorithm only removed two of the 75 principal components and the results were very marginally better. Experimentation with fewer principal component did not yield better results either.

Using PCA and Softmax Regression with keywords as features was attempted as well but the feature set was too large to run successfully. Since CHE categories seemed to be better features using the Multinomial Event Model, it was assumed that this would be the case for Softmax as well, but further experimentation may be warranted.

Additional results

Table 3 has an overview of error rates for some of the other feature combinations which were tried.

Data extracion	Multinomial Event Model													Softmax	
time span	*CHE1*	*CHE2*	*CHE1+2*	*100L1*	*1000L1*	*10000L1*	*50L2*	*100L2*	*1000L2*	*10000L2*	*10L3*	*100L3*	*1000L3*	*CHE1*	*CHE2*
7 days				57%	75%	93%	60%	67%	88%		56%	72%			
15 days	48%	41%			72%	80%			78%	89%			84%	38%	43%
30 days	48%	41%	34%												

***Table 3**: Error rates for some feature combinations that were attempted. CHE1 and CHE2 denotes top level and branch level query categories. nLm denotes top n keywords for all level m advertiser verticals, for example 100L1 uses top 100 keywords (by spend) for each Level 1 (top level) advertiser vertical. Each level has roughly 1.5 orders of magnitude more categories.*

Two takeaways from running feature combinations: (1) Adding more keywords to the mix will produce higher error rates, and (2) Collecting data over a longer time period significantly improves the quality of the Multinomial Event Models with keywords as features, but improves only marginally the Multinomial Event Models with CHE verticals as features (a continuation of this project will further explore the effect on Softmax Regression).

Based upon the first takeaway, additional feature/training sets will be attempted with fewer keywords per vertical (10L1, 50L1).

Future work

As described, additional feature/training sets with fewer keywords per vertical will be explored for the Multinomial Event Model. Further exploration of feature reduction, and reduction of overfitting of the Softmax Regression algorithm will be attempted, as well as exploration of longer data collection periods (30 days).

Exploration of likelihood/confidence cutoff points will be done to increase precision (with low confidence, advertisers will remain unclassified).

The insights produced from this project will be used by Microsoft Bing Ads to classify currently unclassified advertisers.