2015 The 5th International Workshop on Computer Science and Engineering
doi: 10.18178/wcse.2015.04.102

A Cost-Sensitive Ensemble Model for Click-Through Rate Prediction

Hongjian Liu, Defeng Guo

GiantStones Information Technology co., ltd, Shanghai 200032, China

aaron.liu@giantstones.com

Keywords: Click-Through Rate Prediction; Imbalanced data; Cost-Sensitive Ensemble Algorithm; Feature Selection

Abstract. Click-Through Rate prediction is crucial to sponsored search because it can be used to influence ranking, filtering, and pricing of ads. Therefore, estimating click-through rate (CTR) precisely makes significant difference in the efficiency of advertising on the Internet. The CTR prediction can be casted as a binary classification problem (user click as positive class and don't click as negative class) with imbalanced data because the positive class presented with very few samples but associated with a higher identification importance. In this paper, we describe a new cost-sensitive ensemble model for CTR prediction. In this model, we used cost items to denote the uneven identification importance among classes, such that the ensemble strategies can intentionally bias the learning towards classes associated with higher identification importance and eventually improve the identification performance. For feature selection, we extracted two sets of predictive features: basic features and synthetic features. Finally, we made experiments on the dataset of KDD Cup 2012-Track 2 and tested the effectiveness of our model. Experiment results demonstrate that the cost-sensitive ensemble method significantly improve the effectiveness of CTR prediction.

Introduction

Online advertising is a multi-billion dollar industry in the Internet. The online advertising revenue in the full year of 2013 reached over 42.8 billion US dollars [1]. Therefore, choosing the right ads for the query and the order in which they are displayed greatly affects online advertisers, search engine optimizers, and sponsored search providers in the common "pay-per-click" model. To maximize revenue for a search engine, Click-through rate (CTR) is a key indicator since advertisers are charged for each click on their ads, and hence accurate prediction whether a user will click or not click ads is an essential problem for online advertising.

The CTR prediction task can be seen as a regression problem or a binary classification problem. On the one hand, we can use the number of an ad's impressions and the number of the ad's clicks to calculate the click-through rate based on the history information. Then, using the features such as the query topics, the ads descriptions and the user profiles to construct regression function for CTR [2-5]. On another hand, for every ad, which can be represented as 0 (a user does not click the ad) and 1 (the user clicks the ad). Thus, the problem can also be seen as a binary classification [6-10]. According to [10], model the prediction of CTR as a regression problem will results in a regular bias and this approach also lacks of using context features such as user profiles and query topics.

We cast the CTR prediction task as a binary classification problem and define the positive class as these instances that user clicks the ad and the negative class as these instances that user does not click the ad. Literatures have abundant of research based on this approach. Michael et al. [6] built an ensemble of models which combined an artificial neural network and collaborative filters to predict CTR. Graepel et al. [7] proposed a new Bayesian model to predict CTR for Sponsored Search in Microsoft's Bing search engine, which based on a probit regression model that maps discrete or real-valued input features to probabilities. H. Brendan et al. [8] used logistic regression based on an FTRL-Proximal online learning algorithm to predict CTR and explored other issues such as memory savings and performance analysis. Cheng et al. [9] added user-specific and demographic-based features that reflect the click behaviour of individuals and groups to a baseline non-personalized click

model, which significantly improved the prediction accuracy of CTR. Ilya et al. [10] used MatrixNet algorithm to solve the CTR prediction problem while the MatrixNet is the proprietary implementation of boosted trees. As mentioned above, many classification algorithms [6-10] have been well developed and successfully applied to predict the CTR. However, the CTR prediction is an imbalanced data problem which can be characterized as having many more instances of certain classes (a user does not click the ad) than others (a user clicks the ad). According to [11], the imbalance degree of CTR prediction can be as drastic as 1:50, or even larger. Since standard classifiers generally perform poorly on the imbalanced data sets as they are designed to generalize from training data and pay less attention to the rare cases [12-13]. Additionally, noisy data may also make it difficult to learn the imbalance problem.

In this paper, we proposed a cost-sensitive ensemble model for the CTR Prediction. The general idea of the cost-sensitive ensemble approach in dealing with the CTR classification imbalance problem is to boost more weights on the samples in the rare classes, such that the next round of learning will bias towards them. For this purpose, cost items are used for distinguishing different types of samples and the resulting boosting algorithms are regarded as being cost sensitive.

The paper is organized as follows. Section 2 describes the framework and feature selection of our system. Section 3 introduces the cost-sensitive ensemble model, which tackle the task of imbalance problem. Section 4 discuss the experiment design and analyse the experiment results. Finally, we conclude in Section 5.

Click-Through Rate PredictionOrganization of the Text

For an input query, the search engine will retrieve a list of candidate ads. Our target is to predict a user whether click an ad or not for a given query and ad. We formulate this problem as a binary classification problem. First, we collected click and non-click events from search logs as training samples, where each sample represents a query-ad pair presented to a user. In this section, we first provide an overview of our proposed system. Then, we discuss a key step in building the system: feature selection.

System Overview.

As shown in Fig. 1, a search session logs refers to an interaction between a user and the search engine, which is divided into multiple instances, where each instance describes an impressed ad under a certain setting. The related text information was all hash-mapped to integers for privacy. Then, we extract classification features from these raw instances. To train, validate, and ensemble models, we randomly split the dataset into three sets: Training Dataset, Validation Dataset and Test Dataset. In general, our proposed cost-sensitive ensemble model was trained on training dataset and calibrated on validation dataset. Finally, we executed model evaluation on test dataset.

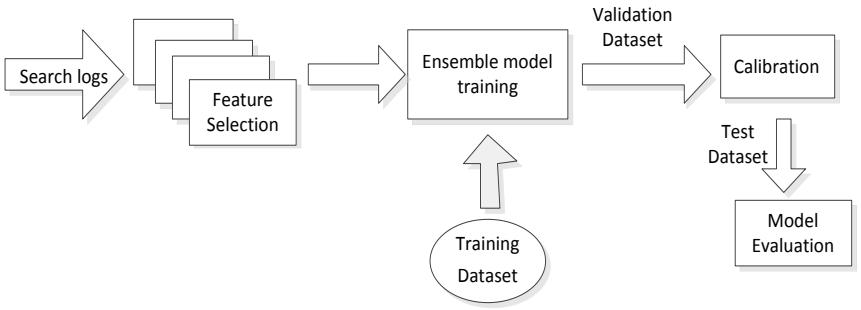

Fig .1: High-level overview of the proposed system

Feature Selection.

Before introducing the cost-sensitive ensemble model for CTR Prediction, we first describe dataset and features used by our experiment here. In this study, we used the dataset of Track 2 of the 2012 KDD Cup competition [14] for experiment. Each instance of a user query and its output was described by 12 fields as presented in Table 1.

Table1: The instance description of dataset

Field member	Description
click	The number of times for a user clicked the ad
Impression	The number of times for an ad displayed to the user
DisplayURL	The shortened landing page URL of the ad
AdID	The ID of an ad
AdvertiserID	The ID of the Advertiser who provide ad
Depth	The number of ads displayed in a search session
Position	The order of an ad in the impression list ads
QueryID	The ID of the user's query in a session
KeywordID	The keywords ID of an ads provided by advertiser
TitleID	The ID of ad's title
DescriptionID	The description ID of an ad
UserID	The ID of a user

In Table 1, for the five ID related fields (QuerylD, KeywordlD, TitlelD, DecriptionlD and UserlD), there are five additional data files to describe the detail information about these fields [14]. Particularly, each line of the QueryID and DescriptionID files maps an id to a list of tokens, corresponding to the query, keyword, ad title, and ad description, respectively. For privacy, each token is represented by its hash value, which means any methods for dealing text can't be used to extract feature.

We extract two kinds of feature vector: one is the basic features and the other is synthetic features.

Basic Features.

As mentioned above, the dataset contains information on UserID, AdID, AdvertiserID, ad's position, etc. We extract basic features directly from the twelve fields of table 1 and the other five additional files. The basic features are described in Table2 and there are 18 features in total. In generally, these features can be divided into three categories: User-related features, Ad-related features and Query-related features.

Table2: The description of basic features for CTR description

Description	**Basic Features**
User-related features	UserID; gender; age
Ad-related features	AdID; position; depth; click; impression; DescriptionID; Description Tokens; TitleID; Title tokens; AdvertiserID; DisplayURL; KeywordID; Keyword Tokens
Query-related features	QueryID; QueryTokens

Synthetic Features.

We observe different group user clustered by the same age or gender may have some similarity that effect the CTR prediction. Fig. 2 shows the average CTR of all users in the dataset. We can see male whose age is between 24 and 30 have the minimum average CTR: 3.73%, while female whose age is more than 40 have the maximum average CTR: 5.2%.

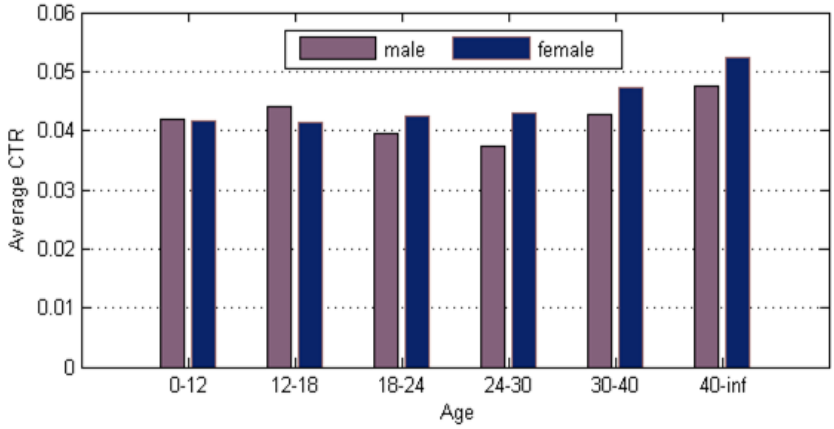

Fig .2: The average CTR of users

In addition, the quality of ads will greatly influence the click behaviour of users, such as the title and description of ads. According to [15], the relative position of an ad also has significant influence on the CTR. Therefore, we compute the average CTR of some categorical features as an additional one-dimensional feature. Take user's gender as an example. For male or female, we compute the average click-through rate for all instances with the same gender, and use this value to synthesis new features， named *gender_CTR*.

For these categories with only a few or even no instances, we use smoothing methods (as shown in Eq.1) to calculate the CTR.

$$\text{smooth} - CTR = \frac{N(\text{click}) + \lambda}{N(\text{impression}) + \gamma} \quad (1)$$

Where λ and γ are smoothing factors.

Finally, the synthetic features used in our experiment are shown in Table 3.

Table 3: The description of synthetic features for CTR prediction

Feature name	Feature Description
Ad_CTR	average click-through rate of AdID
Advertiser_CTR	average click-through rate of AdvertiserID
Depth_CTR	average click-through rate of search session depth
Position_CTR	average click-through rate of ad's position
Title_CTR	average click-through rate of titleID
Description_CTR	average click-through rate of DescriptionID
Keyword_CTR	average click-through rate of KeywordID
Query_CTR	average click-through rate of QueryID
User_CTR	average click-through rate of UserID
Gender_CTR	average click-through rate of user's gender
Age_CTR	average click-through rate of user's age
Ad_Query_CTR	average click-through rate of AdID and QueryID
Ad_User_CTR	average click-through rate of AdID and UserID
User_Query_CTR	average click-through rate of UserID and QueryID

Cost-Sensitive Ensemble Model for CTR Prediction

Ensemble methods was a procedure that combines the outputs of many "weak" classifiers to produce a powerful classifier, which using data modifications at each boosting step consist of applying weights to each of the training observations. Specifically, those observations that were misclassified

by the classifier that induced the previous step have their weights increased at each iteration, whereas the weights are decreased for those that were classified correctly. The AdaBoost algorithm for binary classification [16] is shown in Fig. 3.

1. Initialize the observation weights $w_i = 1/N,\ i = 1, 2, \ldots, N$.
2. For $m = 1$ to M:
   (a) Fit a classifier $G_m(x)$ to the training data using weights w_i.
   (b) Compute
   $$\mathrm{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$
   (c) Compute $\alpha_m = \log((1 - \mathrm{err}_m)/\mathrm{err}_m)$.
   (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))],\ i = 1, 2, \ldots, N$.
3. Output $G(x) = \mathrm{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$.

Fig.3: The Ababoost algorithm for binary classification

For CTR prediction set with imbalanced class distributions, samples of the rare class (user click the ads) are prone to be misclassified. However, the objective of CTR prediction is to improve the identification performance for the rare class. For solving this problem, we draw on the experience of cost-sensitive that boost more weights on those instances associated with higher identification importance. To denote the different identification importance, each instances associated with a cost item: the higher the value, the greater the importance of correctly identifying that instance. Then, the weighting strategy of AdaBoost in figure 3 can be modified as Eq. 2.

$$\omega_i \leftarrow \omega_i . \exp[\mathrm{c}_i . \alpha_i . I(\mathrm{y}_i \neq \mathrm{G}_m(\mathrm{x}_i))],\ i = 1, 2, \ldots, N \qquad (2)$$

where c_i is the cost ratio.

Define the loss function as shown in Eq. 3 using exponential loss function. Thus, the objective is to find α_i minimize the loss function.

$$L(\mathrm{y}, \mathrm{f}(\mathrm{x})) = \exp(-\mathrm{yf}(\mathrm{x})) \qquad (3)$$

For AdaBoost the basic functions are the individual binary classifiers. Using the exponential loss function, one must solve (Eq. 4):

$$(\alpha_m, \mathrm{G}_m) = \arg \min_{\alpha, G} \sum_{i=1}^{N} \exp[-\mathrm{y}_i (\mathrm{f}(\mathrm{x}_i) + \alpha c_i \mathrm{G}(\mathrm{x}_i))] \qquad (4)$$

According to [17], Eq.4 is equal to Eq.5:

$$(\alpha_m, \mathrm{G}_m) = \arg \min_{\alpha, G} \sum_{i=1}^{N} \omega_i^{(\mathrm{m})} \exp[-\alpha c_i \mathrm{y}_i \mathrm{G}(\mathrm{x}_i)] \qquad (5)$$

The solution to Eq.5 can be obtained according to [18] for binary classification:

$$\sum_{i=1}^{N} \omega_i^{(\mathrm{m})} \exp[-\alpha c_i \mathrm{y}_i \mathrm{G}(\mathrm{x}_i)] \leq \sum_{i=1}^{N} \omega_i^{(\mathrm{m})} \left(\frac{1 + c_i \mathrm{y}_i \mathrm{G}_m(\mathrm{x}_i)}{2} e^{-\alpha} + \frac{1 - c_i \mathrm{y}_i \mathrm{G}_m(\mathrm{x}_i)}{2} \mathrm{e}^{\alpha}\right) \qquad (6)$$

Thus, by the first derivation of the right hand side of the inequality (6), α_t can be determined as Eq. 7.

$$\alpha_m = \frac{1}{2} \log \frac{1 + \sum_{i, y_i = f_m(\mathrm{x}_i)} c_i \omega_i^{(\mathrm{m})} - \sum_{i, y_i \neq f_m(\mathrm{x}_i)} c_i \omega_i^{(\mathrm{m})}}{1 - \sum_{i, y_i = f_m(\mathrm{x}_i)} c_i \omega_i^{(\mathrm{m})} + \sum_{i, y_i \neq f_m(\mathrm{x}_i)} c_i \omega_i^{(\mathrm{m})}} \qquad (7)$$

Finally, the cost-sensitive ensemble model proposed in this study for CTR Prediction is shown in Fig. 4.

1. Initialize the observation weights $\omega_i = 1/N, i = 1, 2, \ldots, N$.

2. *For* $m = 1$ *to* M :

(a) Fit a classifier $G_m(\mathrm{x})$ to the training data using weights ω_i.

(b) Compute: $err_m = \dfrac{\sum_{i=1}^{N} \omega_i I(\mathrm{y}_i \neq G_m(\mathrm{x}_i))}{\sum_{i=1}^{N} \omega_i}$

(c) Compute: $\alpha_m = \dfrac{1}{2}\log \dfrac{1 + \sum_{i, y_i = f_m(\mathrm{x}_i)} c_i \omega_i^{(\mathrm{m})} - \sum_{i, y_i \neq f_m(\mathrm{x}_i)} c_i \omega_i^{(\mathrm{m})}}{1 - \sum_{i, y_i = f_m(\mathrm{x}_i)} c_i \omega_i^{(\mathrm{m})} + \sum_{i, y_i \neq f_m(\mathrm{x}_i)} c_i \omega_i^{(\mathrm{m})}}$

(d) Set $\omega_i \leftarrow \omega_i . \exp[\mathrm{c}_i . \alpha_i . I(\mathrm{y}_i \neq \mathrm{G}_m(\mathrm{x}_i))], \; i = 1, 2, \ldots, N$.

3. Output $G(\mathrm{x}) = \mathrm{sign}[\sum_{m=1}^{M} \alpha_m G_m(\mathrm{x})]$

Fig.4: The cost-sensitive ensemble model for CTR prediction

Experiments

In this section, we set up experiments to evaluate the proposed model for CTR prediction. The dataset used in the experiment is Track 2 of the 2012 KDD Cup competition. This is a massive dataset, which includes over 149,639,105 sessions and 22,023,547 users, a total of 10 GB of data. In consideration of the available computing power of our experimental equipment, we randomly selected 737,785 samples for experiment evaluation.

Evaluation Measures. Traditionally, accuracy and recall are the most commonly used evaluation criteria. However, for classification of imbalanced data, these criteria are no longer proper since the rare class has very little impact on the accuracy as compared to that of the prevalent class. Take the CTR prediction for example, the rare class that user click are represented by only about 4% of the training data, a simple strategy can be one that predicts the prevalent class label for every example. It can achieve a high accuracy of 96%. However, this measurement is meaningless to CTR prediction, which aims to identify the rare class. Firstly, we will introduce two measures for evaluating model performance in class imbalance problem.

a) F-measure

F-measure represents a harmonic mean between recall and precision, which is defined as Eq.8.

$$F - measure = \frac{2}{1/p + 1/r} \tag{8}$$

Where p is precision and r is recall.

According to Eq.8, F-measure tends to be closer to the smaller of precision and recall. Hence, a high F-measure value ensures that both recall and precision are reasonably high.

b) G-mean

G-mean measures the balanced performance of a learning algorithm between positive class and negative class. If TP_{rate} represents True Positive Rate and TN_{rate} represents True Negative Rate, G-mean can be defined as Eq.9.

$$G - mean = \sqrt{TP_{rate} \cdot TN_{rate}} \tag{9}$$

Result Analyse. In our proposed cost-sensitive ensemble algorithms, cost ratio are used to boost more weights towards the small class (user click the ads). The cost ratio means the deviation of the learning importance between the two classes. Specifically, it is the ratio of small samples and prevalent samples. In our experiment, we tested a set of cost ratios:[1,2,3,4,5,6,7,8,9,10]. For weak classifier of ensemble learning, it is seldom known in advance which model will perform best for any given problem. However, decision trees is a suitable candidate as an off-the-shelf procedure for CTR

prediction in consideration of many of its advantages, such as natural handling of data of mixed type, insensitive to monotone transformations of inputs and ability to deal with irrelevant inputs and missing values. So, we choose decision trees as the basic classifier for the proposed cost-sensitive ensemble learning model.

The experiment results are shown in Fig. 5. Obviously, the results were sensitive to the cost ratio. When the cost ratio is small, the model was able to achieve higher recall values than precision values with the recall line lying above the F-measure and G-mean line and the precision line below the F-measure and G-mean line. There was an obvious trend that the recall lines descent and precision lines ascent when the cost ratio changing from smaller to larger values.

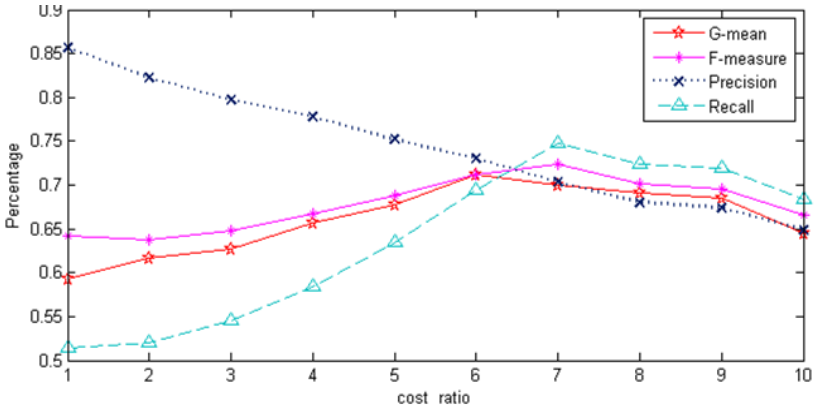

Fig .5: Precision, Recall, F-measure and G-mean with cost ratio = [1,2,3,4,5,6,7,8,9,10] by applying cost-sensitive ensemble model.

In order to further compare the performance of the proposed algorithm, we also applied several traditional classifiers for CTR prediction with the same dataset. As for the AdaBoost algorithm, also choose the decisions trees as the weak classifier. The results are shown in Table 4. Although all these traditional classifiers achieve very high precision values, the recall values are relatively low. Thus, the F-measure and G-mean values are weaker than the cost-sensitive model proposed in this paper.

Table 4: Performance Comparisons among different classifiers

Criterion / Classifier	Precision	Recall	F-measure	G-mean
AdaBoost	0.89	0.51	0.65	0.61
Cost-sensitive model	0.70	0.75	0.72	0.70
SVM	0.92	0.55	0.69	0.53
BayesNet	0.85	0.48	0.61	0.49
Logistic	0.88	0.34	0.49	0.41

Conclusions

In this paper, we proposed a cost-sensitive ensemble model for CTR prediction with the motivation to deal with the imbalance data distribution. The basic idea of this model is to pay more attention and boost more weight to the small class sample. Additionally, we captured various information that user click behaviour and utilized these information to extract two feature sets: basic features and synthetic features. We evaluated the proposed model on a large scale dataset based on logs from the KDD Cup 2012 Track2 competition and observed significant improvements with cost-sensitive ensemble model than those traditional classifiers.

References

[1] Interactive Advertising Bureau, 2013 Interactive Advertising Bureau (2013). IAB Internet Advertising Report-2013 Full-Year Results. Price Waterhouse Coopers. http://www.iab.net/resources/ad_revenue.asp.

[2] Gupta N, Khurana U, Lee T, et al. Optimizing display advertisements based on historic user trails[C]//SIGIR Workshop on Internet Advertising. 2011.

[3] Bauman, K., A. Kornetova, V. Topinskiy, and D. Leshiner. "CTR prediction based on click statistic." Machine Learning in Online Advertising: 8.

[4] Wu K.W., Ferng C. S., Lin H. T., 2012, A Two-Stage Ensemble of Diverse Models for Advertisement Ranking in KDD Cup 2012[C]//KDD Cup Workshop. 2012.

[5] Cheng H, Zwol R, Azimi J, et al. Multimedia features for click prediction of new ads in display advertising[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 777-785.

[6] Jahrer M, Toscher A, Lee J Y, et al. Ensemble of collaborative filtering and feature engineered models for click through rate prediction[C]//KDD Cup Workshop. 2012.

[7] Graepel T, Candela J Q, Borchert T, et al. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine[C]//Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010: 13-20.

[8] McMahan H B, Holt G, Sculley D, et al. Ad click prediction: a view from the trenches[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 1222-1230.

[9] Cheng H, Cantú-Paz E. Personalized click prediction in sponsored search[C]//Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010: 351-360.

[10] Trofimov I, Kornetova A, Topinskiy V. Using boosted trees for click-through rate prediction for sponsored search[C]//Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy. ACM, 2012: 2.

[11] N. V. Chawla, N. Japkowicz, and A. Kolcz. Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explorations Special Issue on Learning from Imbalanced Datasets, 6(1):1–6, 2004.

[12] T. E. Fawcett and F. Provost. Adaptive fraud detection.Data Mining and Knowledge Discovery, 1(3):291–316, 1997.