Knowledge-Based Systems 26 (2012) 246–258

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A probabilistic approach to fraud detection in telecommunications

Dominik Olszewski *

Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland

ARTICLE INFO

Article history:
Received 24 February 2011
Received in revised form 24 August 2011
Accepted 25 August 2011
Available online 3 September 2011

Keywords:
Kullback–Leibler divergence
Latent Dirichlet Allocation
Fraud detection
User profiling
Telecommunications

ABSTRACT

In this paper, a method for telecommunications fraud detection is proposed. The method is based on the user profiling utilizing the Latent Dirichlet Allocation (LDA). Fraudulent behavior is detected with use of a threshold-type classification algorithm, allocating the telecommunication accounts into one of two classes: fraudulent account and non-fraudulent account. The paper provides also a method for automatic threshold computation. The accounts are classified with use of the Kullback–Leibler divergence (KL-divergence). Therefore, we also introduce three methods for approximating the KL-divergence between two LDAs. Finally, the results of experimental study on KL-divergence approximation and fraud detection in telecommunications are reported.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

The Kullback–Leibler divergence (KL-divergence) is a well-known quantity, widely used in probability theory, statistics, and information theory. It was introduced by Kullback and Leibler in [16], and in the probability and statistics context, it evaluates the dissimilarity between two probability distributions, while in the information theory context, it is the measure of relative entropy. The KL-divergence has a wide range of applications, including multivariate data analysis (for example, pattern recognition and discriminant analysis), estimation, approximation, and regression. We focus on its use in fraud detection in telecommunications, which can be regarded as the recognition problem. In our study, this recognition comes down to binary classification, i.e., classification to one of two classes: fraudulent account and non-fraudulent account. We applied a simple threshold-type classification algorithm with automatic threshold setting, which provides computational simplicity and efficiency.

There is a number of fraud detection problems, including credit card frauds, money laundering, computer intrusion, and telecommunications frauds, to name but a few. Among all of them, the fraud detection in telecommunications appears to be one of the most difficult, since there is a large amount of data, that needs to be analyzed, and, simultaneously, there is only a small number of fraudulent calls samples, which could be used as the learning data for the learning-based methods. Consequently, this problem essentially inhibits and limits an application of the learning-based techniques, like the neural-networks-based classifiers, for example.

Fraud detection systems, generally, fall into rule-based systems and user-profiles-based systems. The second of these approaches is regarded as more effective, and became more popular in real-world applications.

1.1. Related work

The general problem of fraud detection has been reviewed in [3,15], while the issue of fraud detection in telecommunications has been studied in [8,4,20,21,10,14,22,13,6]. The authors of [8] present an adaptive and automatic design of user profiling methods for the purpose of fraud detection, using a series of data mining techniques. In paper [20], the Gaussian Mixture Model (GMM) is applied for user profiling, and a high fraud recognition rate is reported. This approach was used in the experimental part of our paper as the comparison for our method. The paper [21] employs Latent Dirichlet Allocation (LDA) to build user profile signatures. The authors assume that any significant unexplainable deviations from the normal activity of an individual user is strongly correlated with fraudulent activity. A straightforward generalization of LDA to time-invariant Markov chains of arbitrary order is proposed in [10], where the experimental study refers to modeling the sequential usage of a telephone service by a large group of individuals. The work [22] presents a novel rough fuzzy set based approach to detect fraud in 3G mobile telecommunication network. A significant contribution in the field of fraud detection in telecommunications belongs to Constantinos Hilas, who investigates the usefulness of applying different learning approaches to a problem of telecommunications fraud detection in

* Tel.: +48 22 234 7618; fax: +48 22 625 6278.
E-mail address: olszewsd@ee.pw.edu.pl

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.08.018

[14], and constructs an expert system, which incorporates both the network administrator's expert knowledge and knowledge derived from the application of data mining techniques on real-world data in [13]. Finally, the recent study [6] aimed at identifying customers' subscription fraud by employing data mining techniques and adopting knowledge discovery process. To this end, a hybrid approach consisting of pre-processing, clustering, and classification phases was applied.

1.2. Our method

Our approach is based on the user profiling technique utilizing LDA, and detecting fraudulent behavior on the basis of threshold-type classification with use of the KL-divergence. Consequently, our method requires the computation of the KL-divergence between two LDAs, which is an unsolved problem. Therefore, this paper focuses also on the issue of approximation of the KL-divergence between two LDAs, introduces three approximation methods, and chooses the most effective one. The fraudulent activity is indicated by crossing a previously defined threshold that causes the fraud alarm. In our paper, also a method for automatic threshold computation is proposed.

The method, proposed in this paper, is based on a classification algorithm that could be applied to any kind of detection problem (not only fraud detection), however, the relation to fraud lies in the fact of using the LDA probabilistic model, and in the fact that the problem of insufficient training data, which is particularly impeding in case of fraud detection, is overcome here. The LDA model provides an accurate description of a user profile, and as it is shown in [21], it can be successfully applied to fraud detection in telecommunications problem. The issue of insufficient training data is overcome, because our method does not require a training process, like it is, for example, in case of the neural-networks-based approaches.

Our technique strongly relies on the user profiling with LDA probabilistic model. Employing LDA for fraud detection in telecommunications was first proposed in [21], however, the difference between [21] and our paper is that we detect whole fraudulent accounts, in contrast to [21], where single fraudulent calls are detected. Consequently, we apply a different classification algorithm with original automatic threshold setting method. This kind of approach is also useful in real-world fraud detection problems. Furthermore, the approach proposed in [21] requires a training phase, while our method is independent of this constraint.

The LDA model is an example of a probabilistic mixture model, i.e., a model described with a combination (linear combination or product) of certain probability distributions. A well-known example of such model is the GMM, being a linear combination of Gaussian distributions. The estimation of LDA model's parameters is described in [2], where the model was introduced, and is described, itself.

Recapitulating, this paper proposes:

- three methods for approximating the KL-divergence between two LDAs,
- a threshold-type classification algorithm for fraud detection in telecommunications,
- a method for automatic threshold computation.

An advantage of our probabilistic approach is that it does not involve the learning process, this way, overcoming associated with it difficulties (insufficient training data).

Furthermore, the proposed method uses only three data sources: the destination, the start-time, and the duration of a call. Consequently, it is not dependent on a large range of variables. This makes the detection easier and faster, and implies numerous benefits in real-world fraud detection systems, for example, the investigator in a large firm does not need to wait for additional data from within the firm, the detection can be done in close to real time, the detection is likely to be cheaper and quicker, because it does not require integration with many other call systems in the firm.

1.3. Remainder of this paper

The rest of this paper is organized as follows: Section 2 presents the KL-divergence and its properties; Section 3 describes the GMM probabilistic model, and reports selected methods for approximating the KL-divergence between two GMMs; Section 4 describes the LDA probabilistic model, and explains, how we employ this model for user profiling; Section 5 introduces the notion of Multinomial Mixture Model (MMM), and proposes two approximation methods of the KL-divergence between two MMMs; Section 6 introduces three methods for approximating the KL-divergence between two LDAs; Section 7 proposes a threshold-type classification algorithm for fraud detection in telecommunications based on the KL-divergence and LDA, and introduces a method for automatic threshold computation; Section 8 reports the results of our experimental study; while Section 9 summarizes the whole paper, and concludes it with some final remarks.

2. KL-divergence (relative entropy)

We have chosen this particular dissimilarity, because of its convenient mathematical form, when it is computed between two probability density functions of a product form, like, e.g., LDAs (see (8) and (16)). The convenience consists in computing a logarithm of the product-functions, and an integral over the probability density function (equal to 1). These advantages of the KL-divergence were used in transformations (12) and (17), which are a part of the main contribution of this paper. There is no other dissimilarity among the well-known quantities in statistics and probability theory providing such properties.

Definition 1 [16,9]. The KL-divergence between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a continuous measurable space Ω is defined as:

$$d(\mathbb{P},\mathbb{Q}) \stackrel{\text{def}}{=} \int_{S(\mathbb{P})} p\log_2\left(\frac{p}{q}\right)\mathrm{d}\lambda, \tag{1}$$

where $S(\mathbb{P})$ is the support of $\mathbb{P}$ on Ω, while p and q are the density functions of measures $\mathbb{P}$ and $\mathbb{Q}$.

Probability measures $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous with respect to the dominating measure λ (for example, λ can be taken to be $(\mathbb{P}+\mathbb{Q})/2$, or can be the Lebesgue measure). Definition 1 is independent of the choice of the dominating measure λ. According to the convention the value $0\log_2\frac{0}{q}$ is assumed as 0 for all real q, and the value $p\log_2\frac{p}{0}$ is assumed as ∞, for all real non-zero p. Therefore, relative entropy takes values from the interval $\langle 0,\infty)$. The KL-divergence is not a metric, since it is not symmetric, and it does not satisfy the triangle inequality. However, it has many useful properties, including additivity over marginals of product measures. If $\mathbb{P}=\mathbb{P}_1\times\mathbb{P}_2$ and $\mathbb{Q}=\mathbb{Q}_1\times\mathbb{Q}_2$ on a product space $\Omega_1\times\Omega_2$,

$$d(\mathbb{P},\mathbb{Q}) = d(\mathbb{P}_1,\mathbb{Q}_1) + d(\mathbb{P}_2,\mathbb{Q}_2). \tag{2}$$

Furthermore, the KL-divergence has the following properties:

1. Self-similarity: $d(\mathbb{P},\mathbb{P})=0$.
2. Self-identification: $d(\mathbb{P},\mathbb{Q})=0$ only if $\mathbb{P}=\mathbb{Q}$.
3. Positivity: $d(\mathbb{P},\mathbb{Q})\geqslant 0$ for all $\mathbb{P},\ \mathbb{Q}$.

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات