CHAPTER 1: INTRODUCTION

1.1 Background

Fraud is a common phenomenon faced by many service providers such as insurance, bank, credit allowance company, computer intrusion, telecommunication etc. Fraud causes a considerable finance impact to all service providers. The Cambridge Advanced Learner defines fraud as “the crime of obtaining money by deceiving people”, while the Concise Oxford Dictionary defines it as a “criminal deception; the use of false representation to gain an unjust advantage”.

As mentioned above, one of the sectors that are particularly vulnerable to fraud case is the telecommunications sector because this sector is a technology-intensive sector. The rapid development of technology and equipment at this time, the development of programming languages, the ease of its use and connectivity have increased the potential incidence of fraud using various new methods that are constantly evolving. On the other hand, the exponential growth of telecommunication data makes the fraud treatment can no longer be handled manually by placing a few dedicated resources. It must be supported by a tool for monitoring the incidence of fraud.

In the telecommunication sectors fraud is defined as any act of cheating, embezzlement in the use of telecommunications facilities intentionally committed by persons or organizations in order to avoid the cost of services or tracking recorded conversation [1]. There are several fraud categories are :

- Subscription fraud is the use of deliberately avoiding the obligation to pay the bills.
- Cloning that user intentionally using telecommunication facilities by charging a usage fee to other customers.
- Bypass fraud that users perform routing of a call without passing official interconnection points.
- Corporate Fraud is a fraudulent act committed intentionally by management, employees, partners or other parties that are fraud, dishonesty, deception and concealment of the truth in order to gain an advantage for the person or party that caused loss to the company or any other party.

One of the services in the telecommunication sectors targeted fraud incident is the service International Direct Dial (IDD) Call. All the IDD service operators faced the same problem, including PT Telkom Indonesia. Although actually, it has many applications based on Voice Over Internet Protocol (VOIP) in the smartphone, laptop or Personal Computer (PC) that can be used to replace these services but the quality of internet broadband in Indonesia

degrades the quality of the service. This led to the use IDD Call is still plenty, especially for events that are considered important which require quality of international call conversation. IDD call service is divided into two categories : using clear channel means use the dedicated international voice network that ensures reliability in quality but has a higher price and using VOIP services that have a lower price and lower quality comparing with clear channel.

IDD Call Service is a service that cannot be built solely by an operator, it involves third parties, called global partner that will integrate voice network operators with international voice networks. These all companies should work together in a contract with a revenue sharing scheme. IDD Call service can be illustrated as Figure 1-1 :

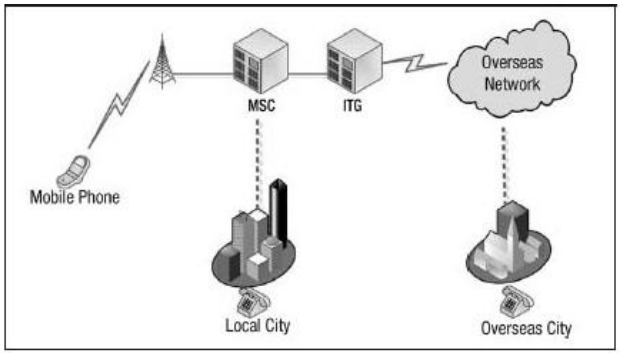

Figure 1-1 IDD Service Diagram

In 2014, PT Telkom Indonesia reported that they had to pay more than 14 billion rupiahs to the global partners because of fraud on international calls. This value is forecast to be continuously increased in the following years if there is no action to address the problem. Fraud is committed by internal customers and Other Licensed Operator (OLO) customers who use its internal call service.

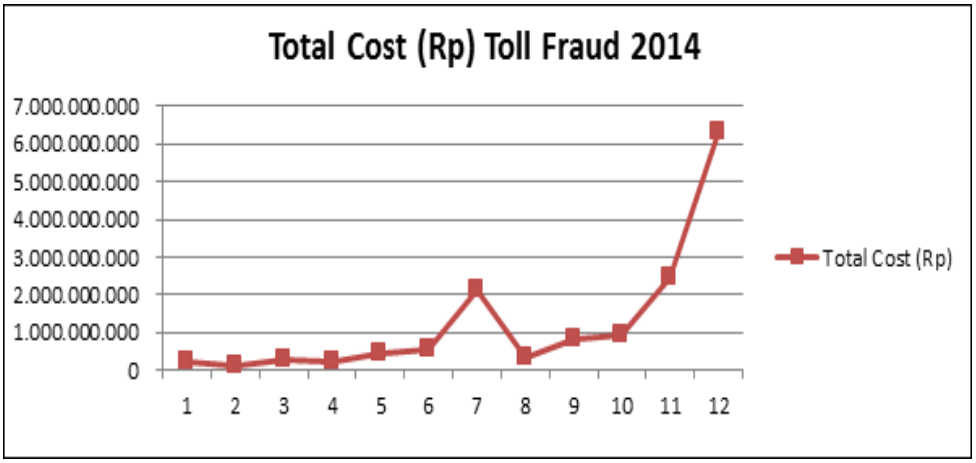

Figure 1-2 Total Cost Toll Fraud 2014

1.2 Theoretical Framework

There have been a lot of research developed for the prevention of fraud. Many data mining methods have been implemented, including classification, clustering, difference measurement or a combination thereof. Latent Dirichlet Allocation (LDA) proved to be applied to the detection of fraud [2], the combination of Latent Derichlet Allocation and Kullback-Leibler Divergence [3], the combination of Support Vector Machine (SVM) and K-Means for clustering followed by Decision Tree for classification [4], Markov Chain and Finite State Automata [5], Neural Network and techniques of probability [6]. Statistical approach [7] and the signature scheme [8] has also been used for the detection of fraud.

Naive Bayesian, Decision Tree and Support Vector Machine (SVM) which included top 10 classification methods [9] is also frequently used in the classification of fraud [10] as well as a hybrid between Naive Bayesian and Kullback-Leibler [11] which will serve as a comparison method in this research.

Kullback-Leibler Divergence [12] give a method to measure divergence or similarity between two probability distributions. Two of probability distributions which have similar information will have zero values of divergence and more dissimilar the divergence value become bigger. This method has proven in applications, e.g. to measure divergence on continuous distribution [13] [14], fraud suspect detection in telecommunication [3] [11], similarity measure on multimedia [15] and calculation of feature weight for feature selection [16].

Latent Derichlet Allocation combined with Kullback-Leibler Divergence to identify fraud suspect [3]. LDA is used as its capability to build a user profiling data [2]. The fraud suspect identified by a threshold of the customer with the fraudulent account and non-fraudulent account. The threshold is calculated using Kullback-Leibler Divergence.

NBTree is a method which combines two algorithms (Decision Tree and Naive Bayesian) [17]. The node in NBTree is a Decision Tree containing a variant of regular decision split, and the leaves are processed as a Naive Bayesian. From another research can be proved that NBTree has a better accuracy in a large database comparing with Decision Tree and Naive Bayesian method. But it has a higher complexity [18] depending on dataset record, class, attribute and mean data per attribute. Another research related to these Decision Tree and Naive Bayesian combination which has the same accuracy with NBTree and a lower complexity is also conducted [19] [20].

NBTree can be used to classify data with several conditions below :

- It has many attributes that are relevant for classification
- The attributes are not necessarily independent
- It can be implemented on a large data

- NBTree is outperform compared with Decision Tree and Naive Bayesian.

At data preprocessing, Principal Component Analysis (PCA) method is used and it is for feature subset solution to determine the dominant feature of the data that will be used in the classification process. PCA is a well-known method used for reducing feature the attribute space and shows the optimal result. The process of using PCA method are : at [21] conducted a comparison of two method feature reduction space : dimensionality reduction and feature subset selection using PCA to see how the influence on classification result, using PCA to train computers with about hundreds of semantic concept with example picture of each concept [22], novel method based on PCA called OR-PCA to process background and foreground object detection in surveillance system which the feature selection used proposed method showed outperformed result comparing with Mixture of Gaussians (MOG), Semi-Soft GoDec (SGD) and Decolor (DEC) [23].

Based on references and research that has been done on Fraud, this research is focusing on the used of Kullback-Liebler Divergence which has proven in the application including on Fraud and will combine with NBTree. The method developed in this research is supervised learning classification which builds a data model based on NBTree algorithms. The classification process is done by calculating the divergence value used Kullback-Leibler divergence formula of data testing and data model.

At the end of the experiment, from the data scenario which provides highest accuracy values will be attempted to be processed using another classification method. Basically, among three data mining methods used in Experiment 1 and Experiment 2 have the same basic method : Naive Bayesian. The Support Vector Machine (SVM) chosen because this method has a different way of classification. Otherwise, SVM is also proven method for fraud detection [15] [24] [25] [26] [27].

1.3 Conceptual Framework

The previous research [11] used Naive Bayesian methods and Kullback-Leibler Divergence to solve fraud problem. This research can solve the misclassification of big customers who have a high duration time of call as a fraudster on previous research [3]. It is also used for the Kullback-Leibler Divergence as a classifier.

In [11] Naive Bayesian is not used independently to perform the classification process due to:

- Naive Bayesian requires a considerable amount of data training to achieve a proper accuracy value. The larger amount of training data, it can achieve higher classification accuracy.

- The amount of fraud data is very small compared to the overall amount of data and it cannot rely solely on the classification method as a single solution.

Therefore, to improve the accuracy and efficiency of the classification, Kullback-Leibler Divergence is used, this is to calculate the amount of divergence between the two probabilities that indicate a significant difference between the normal user and user fraud suspect. It also reduces the need for large data processing.

The previous study [11] algorithm can be described in Figure 1-3 :

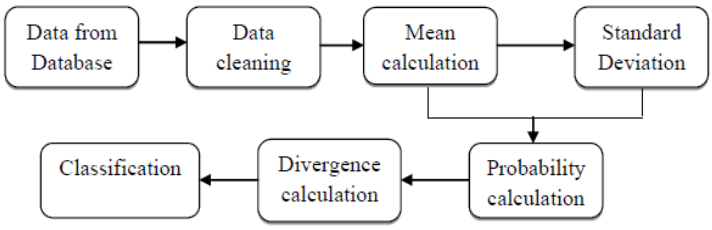

Figure 1-3 Previous studies [11] algorithm using Naive Bayesian – KLD

It can be concluded that the previous study applied a supervised learning classification using Naive Bayesian as a data model builder and Kullback-Leibler Divergence as a classifier by calculating the value of divergence to determine fraudster and non-fraudster customer.

In this thesis, the research is implementing a supervised learning classification method by adding the implementation of the PCA method on preprocessing stage to determine the main dominant attribute against the overall data and NBTree method. This method is to replace Naive Bayesian as a data model builder. The system used in this research can be described as Figure 1-4:

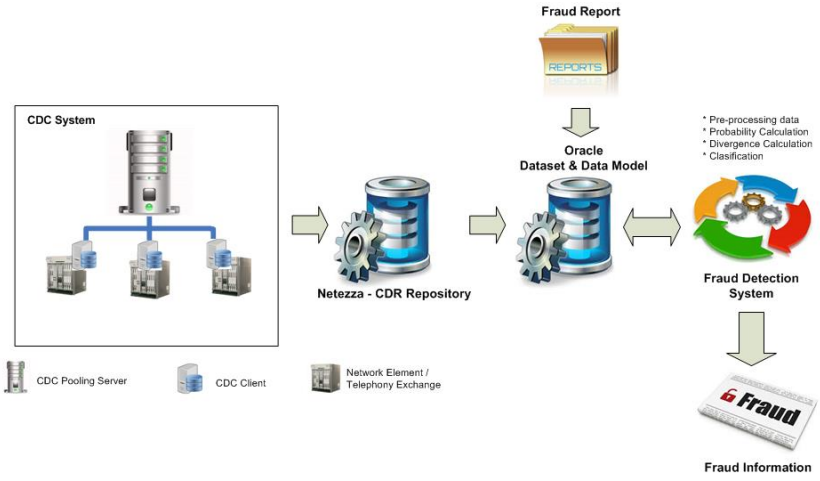

Figure 1-4 Research Framework

The data source is based on International Call Data Record (CDR) provided by PT Telkom Indonesia. The class is specified by the internal fraud report published unit Risk Strategy - Finance Directorate in the same month. Based on Figure 1-4 above, CDR data generated from the Telkom CDC system. The CDR data are already filtered only internal Telkom subscriber (POTS) IDD call traffic by using clear channel service. The total raw data for these purposes are 856.708 call records. On the data preparation stage, the total data used as dataset are 590.595 records which are going to be used on several data scenarios in the experiments.

The entire process is mostly done by the Java language program and Oracle-based database using Oracle Procedural Language/Structures Query Language (PL/SQL).

The proposed method algorithm can be described as the following flowchart :

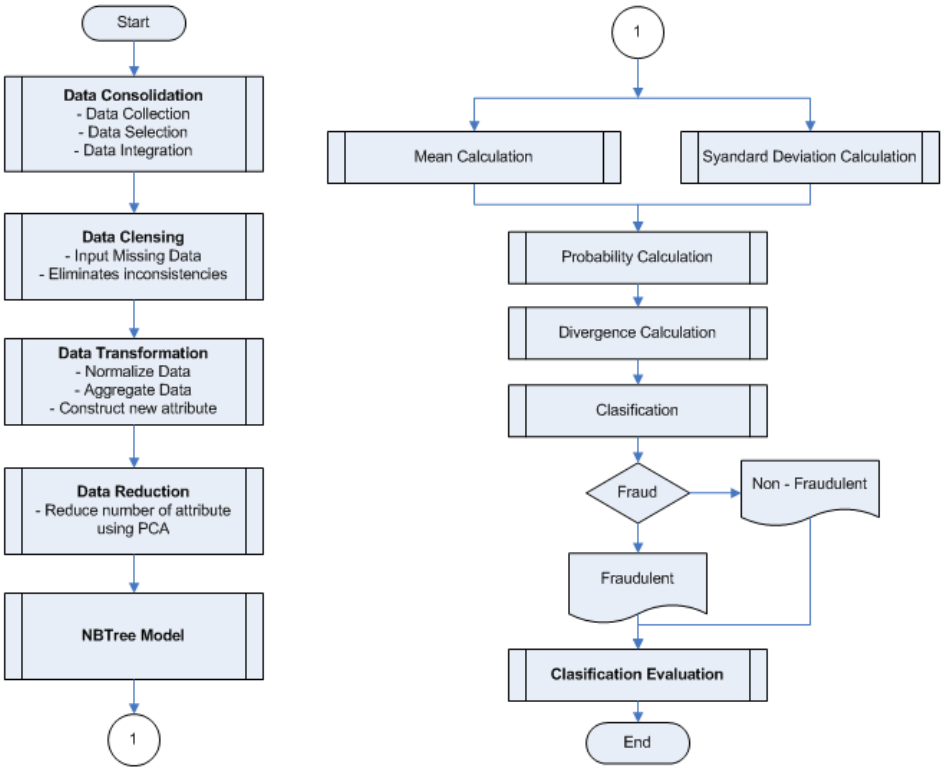

Figure 1-5 Research algorithm

1.4 Objective and Statement of the Problem

Figure 1-2 showed the cost of toll fraud which has to be paid by PT Telkom Indonesia to global partner international call service. The total loss is always increasing and it needs improvement on the existing PT Telkom Indonesia's fraud system detection. The existing system is a simple rule based system which prone of misclassification of fraud that does not comply with the rule. The rule itself can also potentially make misclassification error on the big customer such as a company or person which has many normal international calls. The

data mining method of this research can be used as a new method for fraud detection either as a complement to existing systems by combining them.

The research analyzes a data mining method approach to identified fraud suspects on IDD call service, especially in PT Telkom Indonesia. The main purpose of this research was to minimize the loss because of fraud on IDD call service with increased the accuracy of fraud suspect detection. The method focuses on the used of Kullback Leibler divergence as a classifier. It will be combined with other data mining methods, namely Naive Bayesian and NBTree. Later will be analyzing the improvement the accuracy between those methods. Based on various references, there are some challenging points in fraud suspect detection application in telecommunication sector are [2] [4] [28]:

1. How to improve the accuracy of detection of fraud.
2. How to implement fraud detection on the real environment at a reasonable cost of the process.
3. How to deal with the high volume of call traffic in an efficient manner (large data)
4. How to implement the system on real time data.

The proposed method is expected to be able to overcome the challenges above and work well as a fraud suspect identifier in IDD service in Telkom also in another IDD service provider.

1.5 Hypothesis

Kullback-Leibler Divergence [12] gives a method to measure the divergence or similarity of two probability distributions. The probability of two distributions which have similar information will have zero values of divergence and more dissimilar divergence values become bigger. This method has proven in applications, e.g. to measure divergence on continuous distribution [13] [14], fraud suspect detection in telecommunication [3] [11], measuring similarity on multimedia [15] and calculation of feature weight for feature selection [12]. On research [3], Latent Derichlet Allocation (LDA) combined with Kullback-Leibler Divergence to identify fraud suspect. LDA used as its capability to make a user profiling [2]. The fraud suspect identified by a threshold of the customer with the fraudulent account and non-fraudulent account. The threshold is calculated using Kullback-Leibler Divergence.

C4.5 and Naive Bayes are the top ten data mining algorithm used to create a solution [9] due to their easiness, effectiveness, and efficiency. C4.5 is one of Decision Tree algorithm variants which has an excellent on-time processing. Naive Bayesian has excellent accuracy in classification [29]. NBTree is a method which combines these two algorithms (Decision Tree and Naive Bayesian) [17]. The node in NBTree is a Decision Tree which contains a variant of regular decision split, and the leaves are processed as a Naive Bayesian. From another research

can be proved that NBTree has a better accuracy in a large database comparing with Decision Tree and Naive Bayesian method, but it has a higher complexity [18] depending on dataset record, class, attribute and mean data per attribute. Other research, also conducted related to these Decision Tree and Naive Bayesian combination which has the same accuracy with NBTree and a lower complexity [19] [20].

Some of the advantages possessed by NBTree such as provided better accuracy has a pretty good testing complexity and the ability to handle the data on a large scale. This method is selected to provide a solution of some problems in fraud handling. Additionally, the data were supported by PT Telkom Indonesia in the form of real data IDD Call conversation, this makes this system reliable to be applied in real environmental conditions.

This thesis focuses on combining the advantages of Decision Tree – Naive Bayesian (NBTree) and KL Divergence to gain all its benefit to solving several fraud application problems.

1.6 Assumption

The following is the assumption of this thesis:

a. The experiment is conducted using one month period data, assuming that every month has similar data.
b. International call charge (revenue) is calculated by the average tariff rate of all countries per 60 seconds time unit.

1.7 Scope and Delimitation

To obtain the results of the research so that the expected goals can be achieved, limitations problems of this thesis:

a. Focusing on IDD Call fraud in PT Telkom Indonesia.
b. Focusing on IDD Call with the clear channel service. As described before, the clear channel provides a better quality of service, but it has more expensive charging makes the service become a subscription fraud target.
c. Focusing on IDD Call from Plain Old Telephony Service (POTS) subscriber. All fraud expenses come from POTS subscriber which becomes Telkom responsibility. Otherwise, fraud expense comes from OLO subscriber which becomes OLO responsibility.
d. Due to the limitation and restriction of data access, the data used for the research are international data calls in December 2014 period. This is also based on the reason for the surge in the increases of financial losses by international call fraud as shown in Figure 1-2.

e. Focusing on analyzing the used of Kullback Leibler divergence as a classifier to solve the fraud problem in IDD Call.
f. The study was conducted in the prototype stage.

1.8 Importance of the Study

The ultimate goal of this thesis is to develop a data mining method in the handling of fraud that has a high value of accuracy and better than the previous study. It minimizes the losses suffered by IDD operator caused by fraud, especially in PT Telkom Indonesia. In addition, the method developed using the real data make the method possible to be applied to the real environment with high accuracy with a reasonable cost of the process.

Fraud in IDD service causes a greater impact than fraud on another service because IDD service fraud causes:

- Loss of Revenue:

  Loss of potential revenue from service rendered to the customer because customer rejection of the service has been used. As a result :

  - The revenue can not become the company's earnings.
  - The loss of customer trusts on the service. This can lead to service termination.
- Loss of Gain:

  Operators have to make payments for the sharing service that have been made to the global Partner.