CHAPTER 3

RESEARCH METHODOLOGY

3.1 INTRODUCTION

Nowadays the customers prefer the most accepted payment mode via credit card for the convenient way of online shopping, paying bills in easiest way. At the same time the fraud transaction risks using credit card is a main problem which should be avoided. There are many data mining techniques available to avoid these risks effectively. In existing research they modelled the sequence of operations in credit card transaction processing using a Hidden Markov Model (HMM) and shown how it can be used for the detection of frauds.

Credit card transactions continue to grow in number, taking an ever-larger share of the US payment system and leading to a higher rate of stolen account numbers and subsequent losses by banks. Improved fraud detection thus has become essential to maintain the viability of the US payment system. Banks have used early fraud warning systems for some years. Large-scale data-mining techniques can improve on the state of the art in commercial practice. Scalable techniques to analyse massive amounts of transaction data that efficiently compute fraud detectors in a timely manner is an important problem, especially for e-commerce. Besides scalability and efficiency, the fraud-detection task exhibits technical problems that include skewed distributions of training data and non-uniform cost per error, both of which have not been widely studied in the knowledge-discovery and data mining community.
In today’s increasingly electronic society and with the rapid advances of electronic commerce on the Internet, the use of credit cards for purchases has become convenient and necessary. Credit card transactions have become the de facto standard for Internet and Web based e-commerce. The US government estimates that credit cards accounted for approximately US $13 billion in Internet sales during 1998. This figure is expected to grow rapidly each year. However, the growing number of credit card transactions provides more opportunity for thieves to steal credit card numbers and subsequently commit fraud.

When banks lose money because of credit card fraud, cardholders pay for all of that loss through higher interest rates, higher fees, and reduced benefits. Hence, it is in both the banks’ and the cardholders’ interest to reduce illegitimate use of credit cards by early fraud detection. For many years, the credit card industry has studied computing models for automated detection systems; recently, these models have been the subject of academic research, especially with respect to e-commerce.

The credit card fraud-detection domain presents a number of challenging issues for data mining:

- There is millions of credit card transactions processed each day. Mining such massive amounts of data requires highly efficient techniques that scale.

- The data are highly skewed—many more transactions are legitimate than fraudulent. Typical accuracy-based mining techniques can generate highly accurate fraud detectors by simply predicting that all transactions are legitimate, although this is equivalent to not detecting fraud at all.
• Each transaction record has a different dollar amount and thus has a variable potential loss, rather than a fixed misclassification cost per error type, as is commonly assumed in cost-based mining techniques.

Address the issue of non-uniform cost by developing the appropriate cost model for the credit card fraud domain and biasing our methods toward reducing cost. This cost model determines the desired distribution just mentioned. AdaCost (a cost-sensitive version of AdaBoost) relies on the cost model for updating weights in the training distribution. (For more on AdaCost, see the “Ada-Cost algorithm” sidebar.) Naturally, this cost model also defines the primary evaluation criterion for our techniques.

Furthermore, we investigate techniques to improve the cost performance of a bank’s fraud detector by importing remote classifiers from other banks and combining this remotely learned knowledge with locally stored classifiers. The law and competitive concerns restrict banks from sharing information about their customers with other banks. However, they may share black-box fraud-detection models. Our distributed data-mining approach provides a direct and efficient solution to sharing knowledge without sharing data. We also address possible incompatibility of data schemata among different banks.

Credit card fraud detection has drawn a lot of research interest and a number of techniques, with special emphasis on data mining and neural networks, have been suggested. Ghosh and Reilly have proposed credit card fraud detection with a neural network. They have built a detection system, which is trained on a large sample of labeled credit card account transactions. These transactions contain example fraud cases due to lost cards, stolen cards, application fraud, counterfeit
fraud, mail-order fraud, and non received issue (NRI) fraud. Recently, Syeda et al. have used parallel granular neural networks (PGNNs) for improving the speed of data mining and knowledge discovery process in credit card fraud detection. The complete system has been implemented for this purpose. Stolfo et al suggest a credit card fraud detection system (FDS) using meta learning techniques to learn models of fraudulent credit card transactions. Meta learning is a general strategy that provides a means for combining and integrating a number of separately built classifiers or models. A meta classifier is thus trained on the correlation of the predictions of the base classifiers. The same group has also worked on a cost-based model for fraud and intrusion detection. They use Java agents for Meta learning (JAM), which is a distributed data mining system for credit card fraud detection. Aleskerov et al present CARDWATCH, a database mining system used for credit card fraud detection. The system, based on a neural learning module, provides an interface to a variety of commercial databases. Kim and Kim have identified skewed distribution of data and mix of legitimate and fraudulent transactions as the two main reasons for the complexity of credit card fraud detection. Based on this observation, they use fraud density of real transaction data as a confidence value and generate the weighted fraud score to reduce the number of misdetections. Phua et al suggest the use of meta classifier similar to in fraud detection problems. They consider naïve Bayesian, C4.5, and Back Propagation neural networks as the base classifiers.

3.2 Scope of the Research

The research investigates the hidden markov model problem and discussed its usefulness in the application area such as credit card fraudulent
detection process. Efficient detection of fraudulent behaviour is obtained with the improved precision, recall and F-measure.

3.3 Proposed Methodology

**PHASE 1**

Fraudulent behaviour detection using Semi Hidden with factorized Information Criterion

**SHMM-FIC** → Semi Hidden Markov Model with factorized Information Criterion

**PHASE 2**

Fraudulent behaviour detection using Multiple Hidden Markov Model

**MHMM** → Multiple Hidden Markov Model

**PHASE 3**

Fraudulent behaviour detection using Advanced Hidden Markov Model

**MHMM** → Advanced Hidden Markov Model

**PHASE 4**

Fraudulent behaviour detection using Optimized Semi Hidden Markov Model

**OMSHMM** → Optimized Multiple Semi Hidden Markov Model
Figure 3.1: Research Methodology

Four enhanced algorithms SHMM-FIC, MHMM, AHMM, OMSHMM are proposed in this research to provide a potential solution for IDS problem.

The figure 3.1 shows the research methodology of this proposed approach. It consists of mainly four phases such as:

1. Fraud Behaviour detection using Semi Hidden Markov Model with Factorized Integrity Checking

2. Fraud behaviour detection using Multiple Hidden Markov Model

3. Fraud behaviour detection using Advanced Hidden Markov Model

4. Fraud behaviour detection using Optimized Multiple Semi Hidden Markov Model

3.3.1. Fraud Behaviour detection using Semi Hidden Markov Model with Factorized Integrity Checking

To provide better accuracy and to avoid computational complexity in fraud detection in proposed work semi Hidden Markov model (SHMM) algorithm of anomaly detection is presented which computes the distance between the processes monitored by credit card detection system and the perfect normal processes. With this we are implementing another method for fraud detection is that having a key idea is to factorize marginal log-likelihood using a variation distribution over latent variables.

An asymptotic approximation, a factorized information criterion (FIC) obtained by applying the Laplace method to each of the factorized components. This
method is also having a larger class of practical problems that can be properly modelled in the setting of SHMM. Also major constraint is found, conversely, in mutually HMM and SHMM, i.e., it is generally imagined that there survives at least one observation connected with every state that the hidden Markov chain takes on.

### 3.3.2. Fraud behaviour detection using Multiple Hidden Markov Model

To improve the efficiency of SHMM in our proposed research we are combining the multiple observation of SHMM called Multiple Semi Hidden Markov Model (MSHMM) through this we can improve the detection accuracy better than the SHMM. In normal Hidden Markov Model the problem of cannot find an optimal state sequence for the underlying Markov process also this observed sequence cannot be viewed as training a model to best fit the observed data.

### 3.3.3. Fraud behaviour detection using Advanced Hidden Markov Model

In this research, the main aim is to model the sequence of observations in credit card transaction processing using an Advanced Hidden Markov Model (AHMM) and show how it can be utilized for the exposure of frauds. In this process an AHMM is initially trained with the regular manners of a cardholder. If an incoming credit card transaction is not recognized by the trained AHMM with adequately high probability, it is believed to be fraudulent. This proposed work desire to regulate the model parameters to best fit the observations.

The ranges of the matrices (N and M) are fixed but the elements of A, B and \( \pi \) are to be decided, focus to the rank stochastic condition. The information that can efficiently re-estimate the model itself is one of the more incredible features of HMMs this referred here as AHMM. The experimental results shows that we can
significantly reduce loss due to fraud through distributed data mining of fraud models through this proposed work of AHMM.

3.3.4. Fraud behaviour detection using optimized Multiple Semi Hidden Markov Model

Finally, optimized Multiple Semi Hidden markov chain model is proposed in this research to optimize the results that are obtained in all its phases. This is achieved by selecting an optimal model training parameter for the effective prediction of fraudulent behaviour. The cuckoo search algorithm is used in this work for optimized selection of model training parameter. Cuckoo search (CS) is an optimization algorithm developed by Xin-she Yang and Suash Deb in 2009. It was inspired by the obligate brood parasitism of some cuckoo species by laying their eggs in the nests of other host birds (of other species). Some host birds can engage direct conflict with the intruding cuckoos. For example, if a host bird discovers the eggs are not their own, it will either throw these alien eggs away or simply abandon its nest and build a new nest elsewhere. Some cuckoo species such as the New World brood-parasitic Tapera have evolved in such a way that female parasitic cuckoos are often very specialized in the mimicry in colours and pattern of the eggs of a few chosen host species.

Cuckoo search idealized such breeding behaviour, and thus can be applied for various optimization problems. It seems that it can outperform other metaheuristic algorithms in applications.

Cuckoo search (CS) uses the following representations:
Each egg in a nest represents a solution, and a cuckoo egg represents a new solution. The aim is to use the new and potentially better solutions (cuckoos) to replace a not-so-good solution in the nests. In the simplest form, each nest has one egg. The algorithm can be extended to more complicated cases in which each nest has multiple eggs representing a set of solutions.

**CS is based on three idealized rules:**

1. Each cuckoo lays one egg at a time, and dumps its egg in a randomly chosen nest;

2. The best nests with high quality of eggs will carry over to the next generation;

3. The number of available host’s nests is fixed, and the egg laid by a cuckoo is discovered by the host bird with a probability $p_d \in (0, 1)$. Discovering operates on some set of worst nests, and discovered solutions dumped from farther calculations.

**3.4. SUMMARY**

In this chapter we discussed about overview of thesis, background details of Hidden Markov Model (HMM) and how this HMM can be applied to detect credit card fraudulent activity. In forthcoming chapter we deal this methodology in detail.