Accuracy Evaluation of a Credit Card Fraud Detection System on Hadoop MapReduce

Elham Hormozi
Computer Engineering and Information Technology Mazandaran University of Science and Technology,
Babol, IRAN
e.hormozi@ustmb.ac.ir

Hadi Hormozi
Computer Engineering and Information Technology Arak University,
Arak, IRAN
h.hormozi@qazd.ir

Mohammad Kazem Akbari
Computer Engineering and Information Technology Amirkabir University of Technology (Tehran Polytechnic) Tehran, IRAN
akbarif@aut.ac.ir

Morteza Sargolzaei Javan
Computer Engineering and Information Technology Amirkabir University of Technology (Tehran Polytechnic) Tehran, IRAN
msjavan@aut.ac.ir

Abstract—**In digitalization era, credit card fraud detection is of high significance to financial organizations. This paper discussed about credit card fraud detection by parallelizing of Negative Selection Algorithm on the Cloud computing platform. We present performance evaluation of running the algorithm on the cloud by MapReduce framework and show it's dramatically results on real world financial data. We argue that, for the fraud detection rate, False Negative rate, fraud catching rate (True Positive rate) and false alarm rate (False Positive rate), Cost and Hit rate that are the best metrics for a desirable credit card fraud detection system.**

Keywords-Credit card fraud detection; Parallelize; Negative Selection Algorithm; Cloud computing; MapReduce.

I. INTRODUCTION

These days, financial institutions usually develop custom fraud detection systems targeted to their own asset bases. Novelty banks have come to realize that a global approach is required, involving the periodic sharing with each other information about frauds. Such information sharing is the basis of building a global fraud detection infrastructure where local detection systems propagate attack information to each other. The key difficulties in building such a system are: 1. financial companies don't share their data for a number of (competitive and legal) reasons. 2. The databases that companies maintain on transaction behavior are huge and growing rapidly, which demand scalable machine learning systems. 3. Real-time analysis is highly desirable to update models when new events are detected. 4. Easy distribution of models in a networked environment is necessary to keep up to date detection capability [1, 2]. We have proposed a novel system to address these issues. Our system is based Artificial Immune System [12]. We have used the Negative Selection Algorithm [9, 10, 11] and parallel it by the MapReduce [14] on Apache Hadoop platform [16, 17].

By using of cloud computing can access to share database between organizations and also the cloud systems have a high computing power and other advantages.

The structure of this paper is as follows: first we introduce the reader to credit card fraud detection. In Section III, we briefly explain the Artificial Immune System and Negative Selection Algorithm. In Section IV, we introduce cloud computing, hadoop and mapreduce model. In Section V, we present our propose solution. In Section VI, we discussed about gained results of experiments. This paper finished by a conclusion.

II. CREDIT CARD FRAUD DETECTION

Credit cards are a fine destination for fraud, since in a very short time a large amount of money can be earned without taking too many hazards. This is because frequently the crime is only detected several weeks after date [1]. Credit card fraud is the abusage of a credit card to make purchases without permission or counterfeiting a credit card [3]. Types of credit card fraud are: online credit card fraud, shave and paste, stolen card numbers, advance payments and etc. If current growth rates continue, credit cards and debit cards will each exceed the number of paid checks before the end of the decade. As the industry continues to expand and offer credit to more and more consumers, fraud will also grow Fraud detection is, given a set of credit card transactions, the process of identifying those transactions that are fraudulent. A desirable fraud detection system need to good metrics to evaluate the system. The system should take into account the cost of the fraudulent behavior detected and the cost associated with stopping it. In fact, there is a decision layer on top of the fraud detection system. This layer decides what actions to take when fraudulent behavior is detected via the fraud detection system [1, 4].

978-1-4673-6490-4/13/$31.00 ©2013 IEEE

III. ARTIFICIAL IMMUNE SYSTEM

Wide spectrum of security issues environs e-commerce. For instance, a big problem encountered by e-commerce retailers is the problem of online credit card payment fraud. Essentially, the online credit card fraud detection problem range is a classification problem which transactions need to be classified as either legal or abnormal. Many of systems are fighting to confront with complex and quickly evolving areas like this. Based on more comparative solutions have been follow [12]. The biological immune system discriminates between self-proteins and the foreign antigens that compose non-self. It is a natural classifier. Because of that, artificial immune system (AIS) is an interesting method to online credit card fraud detection. The biological immune system is composed of a diverse domain of cells, but of special interest to most AISs are the cells known as lymphocytes [5]. White blood cells (Lymphocytes) are the immune system's antigen detectors, where antigens are detrimental foreign organisms. Lymphocytes contain receptors on their area, when the affinity between the receptor and the antigen is adequate permit them to detect antigens [6, 7, 8]. Figure **1** shows antigens at the low left that attach to the lymphocyte with high affinity.

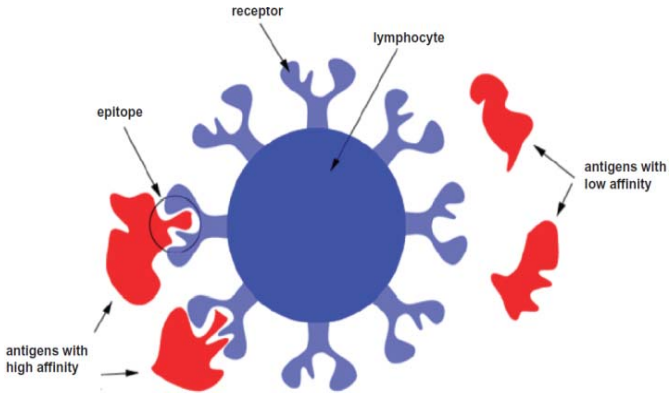

Figure 1. Binding of a lymphocyte and foreign antigens [8]

A. Negative Selection Process

Unlike antigens, self-proteins are generated by the human body and they are beneficial organisms [9]. In order to prevent the detectors attaching to self-cells, immature lymphocytes bearing a process that called Negative Selection (Figure **2**). Pending this process, if a lymphocyte attaches to a self-cell, the lymphocyte will be death. If the lymphocyte survives being exposed to organisms for an enough time period, the lymphocyte becomes a mature operational lymphocyte [10, 11].

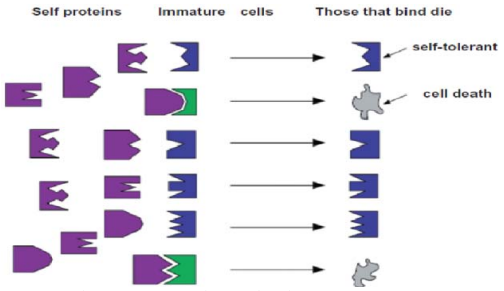

Figure 2. Negative Selection Process [10]

B. The Negative Selection Algorithm

The main idea of this algorithm is to generate a set of detectors by first randomly making candidates and then discarding those that recognize training self-data, and then these detectors can later be used to detect anomaly [13]. The starting point of this algorithm is to produce a set of self strings that define the normal state of the system. The task then is to produce a set of detectors that only recognize the complement of self-cells. These detectors can then be applied to new data in order to classify them as being self or non-self.

1. *Steps to build the Fraud Detection System*
   - b) *Case library initialization (normal transactions and fraud transactions case)With supervised classification*
   - c) *Antibody Gene (detectors.)library Initialization - Randomly*
   - d) *Negative selection of detectors*
2. *Fraud Detection*

   The fraud detection function is activated, when a new transaction is submitted for fraud detection. The affinity between the antibodies (detectors) in the gene library and the new antigens is calculated. If the affinity threshold set by the system is exceeded, the AISFD[1] system sends out a fraud alert.

IV. CLOUD COMPUTING

Cloud computing refers to an architecture where the user accesses software using the Internet. In order to receive these services, users only need sufficient infrastructure to connect to the seller's servers. Users needn't know about the internal workings of the software or the location of the servers [14]. Cloud computing is a scalable technology rend for developing world adoption, helping lower costs, expand operation flexibility and improve speed of service. Most cloud sellers charge by using of resources in processing hours, gigabits consumed, and gigabits per second transferred, rather than by monthly costs. Cloud computing is so powerful and very accessible. Connecting hundreds or thousands of computers together in a cloud creates riches of computing power impossible with a single desktop PC. Also with cloud systems there is the ability to share a database including of several types of frauds between multiple organizations [15].

A. Hadoop Mapreduce

Hadoop MapReduce is a software framework for easily writing applications which process huge quantitise of data in-parallel on massive clusters of hardware in a reliable and fault-tolerant method. As you seen in Figure 3, a MapReduce job usually scatters the input dataset into independent chunks which are processed by the map tasks in a completely parallel way. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Usually both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-

[1] **A**rtificial **I**mmune **S**ystem **F**raud **D**etection

executes the failed tasks. The MapReduce framework include of a single master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master. This advantages resolve the problems of paralleling of tasks [21, 22].

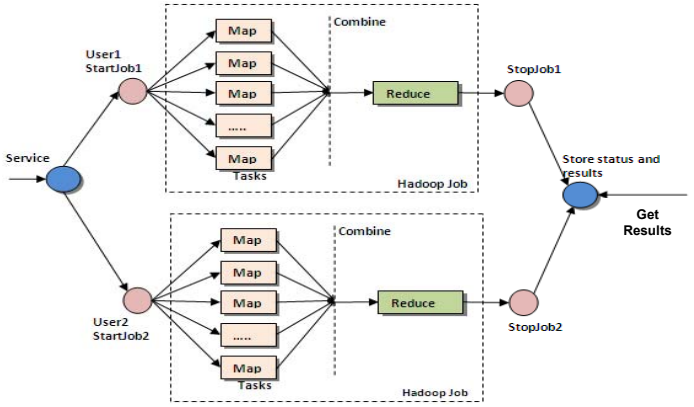

Figure 3. Hadoop MapReduce Structure [16]

V. OUR PROPOSED MODEL

Architecture of the final model according to previous sections has shown in Figure **4**. Our proposed fraud detection model using of cloud platform for detecting the several types of credit card frauds.

This model with having data from different organizations can be applied the fraud detection algorithm on them and extract fraud models. We parallel NSA by mapreduce programming model on apache hadoop. This means, the detectors generate by mapreduce mechanism. Two functions that called Map and Reduce add to NSA at the first time.

The NSA consists of two phases, Training phase and Test phase. In the first phase of the algorithm, data should be prepared and normalized. So, detectors produced randomly. In this phase by using of map function Euclidean distance between any two records is calculated in training set and the threshold can be determined. So, by using of the reduce function the average distance is calculated and output is recorded. Afterwards, start up the main phase of training and producing the detectors using of map function. Map function are divided the input data among the reduce functions then detectors is produced in the map function and finally registered to output. Then, the affinity threshold between the detector and the train set of dataset is calculated by use of Euclidean distance too. The other distance measurements such as Manhattan and Hamming can be use for affinity threshold distance. If the distance be less than the threshold, it means that the detector detect a self record, otherwise it, detector recognized a non self record. As mentioned, only training phase was run in parallel and the test phase was perform in serial like the basic NSA.

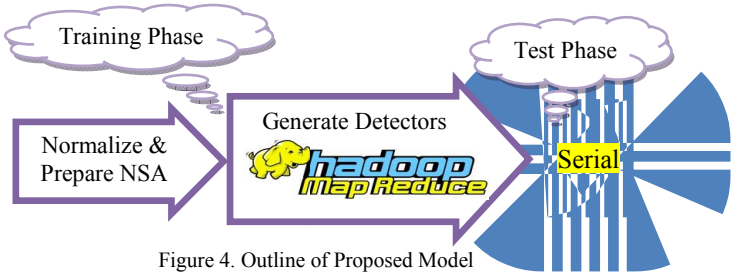

Figure 4. Outline of Proposed Model

A. DataSet

We have obtained a large dataset, 300000 records, of credit card transactions from a large Brazilian bank, with registers within time window between Jul/14/2004 through Sep/12/2004. All data fields are considered in numerical form. The dataset consist of 17 fields. Each of field is a column. Each of a row in dataset called a record. We split dataset into two parts. Part1 is including of 70 percent of dataset and assigned to trainset. Part2 is including of 30 percent of dataset that allocated to testset.

B. Metrics of Evaluation

The overall accuracy is momentous. But for the fraud detection area the fraud catching rate (True positive), the false negative rate, detection rate and cost are the critical metrics. A low fraud catching rate means that a great of forensic transactions will go through our detection. Totally, the stream of accounts flagged by the fraud detection system includes the compromised accounts, true positives (TP), and incorrectly implicated cases, false positives (FP). Complementarily, accounts that pass through the fraud detection system with no alert created are a mixture of legitimate accounts, true negatives (TN) and missed fraudulent cases, false negatives (FN). Several performance criteria can be used in an application to the fraud detection problem. It is typically considered that the error committed in assessing a fraudulent case as legitimate (FN) is more serious than the complementary type of error (FP) [18].

VI. EXPERIMENTS

We considered two amount of detector number for running the paralleled NSA on the cloud environment. Also, we executed the algorithm once in serial and once in parallel with 5 different types of mapper values. The formulas that used for evaluation present in Table **1**. Also the result of our experiments presented in Table **2** and Table **3**.

A. Formula

The formulas that used in this paper are been applied in many papers.

Table 1. Used Formulas

1	$\text{Normalization} = \frac{\text{Value(field)} - \text{Min(field)}}{\text{Max(field)} - \text{Min(field)}}$
2	$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n}(V1_i - V2_i)^2}$
3	False Negative Rate = FN / (FN + TP) [20]

4	False Positive Rate = FP / (FP+TN) [19]
5	True Positive Rate= TP/(TP+FN) [20]
6	True Negative Rate= TN/(TN+ FP) [19]
7	Detection Rate = TP / (TP+FN) [20]
8	Hit Rate = TP / (TP+FP) [19]
9	$Cost = $100 x FN + $10 x FP + $1 x TP [19]

B. Results

We apply two detector numbers for running NSA on the cloud and compare the gained results of them with each other. The results of executing the algorithm with detectors number= 100000 have been shown in Table **2**.

Table 2. Results of running NSA on hadoop with DN=100000

Detector Number= 100000					
Negative Selection	Mapper	TN	TP	FN	FP
Parallel	2	108487	2401	1145	84
	4	108287	2436	1132	262
	8	107755	2945	1245	172
	16	106326	4129	1161	501
	24	106061	4605	858	593
Serial		108527	1390	2199	0

The results of NSA on the cloud with several mappers for detector number= 200000 have been display in Table **3**.

Table 3. Results of running NSA on hadoop with DN=200000

Detector Number= 200000					
Negative Selection	Mapper	TN	TP	FN	FP
Parallel	2	108569	2510	998	40
	4	108264	2769	1009	75
	8	107919	2988	981	229
	16	107220	868	894	708
	24	76293	3295	341	701
Serial		108577	1772	1768	0

According to the obtained results of Table **2** and Table **3** and using the results in the formulas of Table **1** we evaluated them and compared the results of parallel algorithm on cloud with multiple types of mappers. As you seen in Table **4**, false negative rate in the basic algorithm is around 61 percent, while with paralleled NSA on the cloud and execute with several types of mapper this rate reduced to 15 percent. Also the true positive or detection rate increased from 38 percent to 84 percent approximately. Based on assessments was conducted the false positive rate and true negative rate haven't been good in practice. But, false negative rate and true positive are more important than the false positive rate and true negative rate for financial institutes [18]. Also the hit rate a little decreased that in comparison to reduce cost is negligible. As you can see, the cost of organizations decreased from 221290$ to 96335$.

Table 4. The final results with Detector Number= 100000

Detector Number	100000					
Algorithm	Seria Negative Selection	Parallel Negative Selection on the Cloud				
Mappers Number		2	4	8	16	24
False Negative Rate	61.27%	32.29%	31.73%	29.71%	21.95%	15.71%
False Positive Rate	0.00%	0.08%	0.24%	0.16%	0.47%	0.56%
True Positive Rate	38.73%	67.71%	68.27%	70.29%	78.05%	84.29%
True Negative Rate	100.00%	99.92%	99.76%	99.84%	89.18%	88.59%
Detection Rate	38.73%	67.71%	68.27%	70.29%	78.05%	84.29%
Cost	221290	117741	118256	129165	125239	96335
Hit Rate	100.00%	96.62%	90.29%	94.48%	89.18%	88.59%

As you can see in Table 5, increasing the number of detector number will increase the accuracy. The results of below table are better than compare to Table **4**.

Table 5. The final results with Detector Number= 200000

Detector Number	200000					
Algorithm	Seria Negative Selection	Parallel Negative Selection on the Cloud				
Mappers Number		2	4	8	16	24
False Negative Rate	49.94%	28.45%	26.71%	24.72%	21.34%	6.92%
False Positive Rate	0.00%	0.04%	0.07%	0.21%	0.66%	0.65%
True Positive Rate	50.06%	71.55%	73.29%	75.28%	28.66%	93.08%
True Negative Rate	100.00%	98.43%	97.36%	92.88%	82.31%	86.74%
Detection Rate	50.06%	71.55%	73.29%	75.28%	78.66%	93.08%
Cost	178572	102710	104419	103378	99775	45695
Hit Rate	100.00%	98.43%	97.36%	92.88%	82.31%	86.74%

CONCLUSION

Because of the credit card fraud detection in financial institutes is an important subject; we decided using of AIS into cloud for having a good fraud detection system. Also, we did accuracy evaluation for it. Cloud computing have a powerful computing power and provided a shared database of several frauds between many of banks and organizations. Cost, false

negative rate, detection rate, true positive rate and so on, are major criteria for a desirable fraud detection system. Our experiments done on hadoop mapreduce platform. Also, we perform the changes on NSA and used several types of mappers. Two detector set was used for the experiments. For example, the cost reduce about 74 percent toward serial NSA for detector number= 200000 and it decrease around 56 percent for detector number= 100000.

ACKNOWLEDGMENT

We would like to thank all those who contributed to this paper. Further to this, we gratefully acknowledge those in the cloud computing team at the Department of Computer engineering and Information Technology, Amirkabir University, IRAN.

REFERENCES

[1] Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B. Credit Card Fraud Detection using Bayesian and Neural Networks. Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies, (2002).

[2] Salvatore J. Stolfo, David W. Fan, Wenke Lee, Andreas Prodromidis and Phil K. Chan. 1997. Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results. Technical Report CUCS-008-97, Computer Science Department, Columbia University.

[3] Balan, Lăcrămioara, and Mihai Popescu. "Credit card fraud." The USV Annals of Economics and Public Administration 11.1 (2011): 81-85.

[4] Tetro, Donald, Edward Lipton, and Andrew Sackheim. "System and method for enhanced fraud detection in automated electronic credit card processing." U.S. Patent No. 6,095,413. 1 Aug. 2000.

[5] J. Hunt and D.Cooke. "Learning Using an Artificial Immune System," journal of network and computer applications. 19:189-212,1996.

[6] De Castro, L.N. & Von Zuben, F.J. (1999a) Artificial immune systems: part i – basic theory and applications. Technical Report, Department of Computer Engineering and Industrial Automation, School of Electrical and Computer Engineering, State University of Campinas, São Paulo, Brazil, December.

[7] De Castro, L.N. & Von Zuben, F.J. (1999b) Artificial immune systems: part ii – a survey of applications. Technical Report, Department of Computer Engineering and Industrial Automation, School of Electrical and Computer Engineering, State University of Campinas, São Paulo, Brazil, December.

[8] Tuo, J., Ren, S., Liu, W., Li, X., Li, B. & Lei, L. (2004) Artificial immune system for fraud detection. 2004 IEEE International Conference on Systems, Man and Cybernetics, October 2004, v. 2, pp. 1407–1411.

[9] Forrest, S., Perelson, A.S., Llen, L. & Cherukuri, R. (1994) Self-nonself discrimination in a computer. Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy, May, pp. 202–212.

[10] Forrest, S., Hofmeyr, S.A. & Somayaji, A. (1997) Computer immunology. Communications of the ACM, 40, 88–96.

[11] Wightman, J. (2003) Computer immune techniques in e-commerce fraud detection. School of Information Systems and Technology Management, The University of New South Wales, Honours Thesis, 2003.

[12] Nicholas Wong, Pradeep Ray, Greg Stephens, Dr. Lundy Lewis: "Artificial Immune Systems for the Detection of Credit Card Fraud: An Architecture, Prototype and Preliminary Results", Information Systems Journal, vol 22 issue 1, pp. 53-76, Jan. 2012.

[13] E. Hormozi, M. K. Akbari, M. S. Javan, H. Hormozi,"Performance Evaluation of Fraud Detection based Artificial Immune System on the Cloud," In procciding of the 8th International Conference on Computer Science & Education, Colombo, Seri Lanka IEEE, 2011.

[14] Staten, J., Yates, S., Gillett, F. E., Saleh, W. and Dines, R.A. (2008). Is cloud computing ready for the enterprise? Forrester Research, Cambridge, MA.

[15] M. Miller, Cloud Computing: Web-Based Applications That Change the Way You Work and Collaborate Online: Que, Aug. 2008.

[16] http://hadoop.apache.org/hdfs/docs/current/hdfs_design.html visited on Feb. 9th 2012.

[17] T. White, Hadoop: The Definitive Guide, O'Reilly, 2nd Edition, pp. 15-32, 2011.

[18] M. Krivko, "A Hybrid Model For Plastic Card Fraud Detection Systems", Expert Systems with Applications, Vol. 37, No. 8, pp. 6070-6076, 2010.

[19] A. Brabazon, et. al., "Identifying Online Credit Card Fraud using Artificial Immune Systems", IEEE Congress on Evolutionary Computation (CEC), Spain, 2011.

[20] N. Wong, et al., "Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results", Information Systems Journal, Vol. 22, No. 1, pp. 53–76, 2012.

[21] J. Lin and C. Dyer. Data-Intensive Text Processing with MapReduce. Morgan & Claypool Publishers, 2010.

[22] T. White, Hadoop: The Definitive Guide. O'Reilly Press, second ed., 2010.