International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

An Efficient Approach for Privacy Preserving Data Mining using SMC Techniques and Related Algorithms

P Annan Naidu[1], Dr. M. Vamsi Krishna[2]
[1, 2]*CSE, Centurion University, Odisha, India.*

Abstract: ***Data mining plays a major role in real world business applications by providing different techniques and algorithms. When data extracted from different data sources for business decisions or for business processing, It is mandatory to secure the data of individuals or group and providing privacy of data. When the information shared among different nodes such that centralized or distributed , then data mining results should ensure the secret sharing of information, This paper presents secure multi-party computations for privacy preserving data mining and also states that different approaches to achieve secure multi-party computations, in this process SMC deals with secret sharing of data among different nodes or parties. This paper describes about SMC role and its techniques along with algorithms.***

Keywords: Data Mining, distributed Data Mining, Randomization, Anonymization, Secure Multy-Party Computation

I. INTRODUCTION

Data mining is the process of extracting knowledge. Data mining is one of the emerging technologies for data analysis and classification using different clustering algorithms. Past research in data mining discussed many security issues, solutions with corresponding methods and tools. In privacy preserving data mining (PPDM), the aim is to carry out data mining operations on datasets without revealing the contents of the private data. Since the outputs of the mining notify us something regarding the data, a few information about the real data is revealed to the mining results. This directs to the loss of privacy. If the data is disturbed on the other side for privacy fears, it directs to data loss, which usually refers to the quantity of essential information preserved about the datasets after the perturbation [3].

In Data Distribution, Data is distributed across many sites, in order to get the original data, data must be combined from all sites. Data distribution is used as a privacy preserving scheme, Where the original data is distributed among sites, no site knows the data of other site, so data recollection cannot be done through any single party. Distributed computing plays an important role in the Data Mining process for several reasons. First, Data Mining often requires huge amounts of resources in storage space and computation time. To make systems scalable, it is important to develop mechanisms that distribute the work load among several sites in a flexible way. Second, data is often inherently distributed into several databases, making a centralized processing of this data very inefficient and prone to security risks. Distributed Data Mining explores techniques of how to apply Data Mining in a non-centralized way.[20]

Privacy preserving distributed data mining is the extraction of relevant knowledge from large amount of data, while protecting at the same time sensitive information or personally identifiable information in the unstructured distributed environment [2]. The most important example of a distributed data environment is Internet where data storage and processing increasing rapidly that deal with several areas. In Distributed data mining, the entire data may be sited at a single location or distributed at various sites referred to as centralized or distributed respectively. In a distributed environment, multiple sites logically interrelated data are distributed among various sites. The distribution of data may be horizontal, vertical or hybrid. In horizontal partitioning each site contains a subset of records of the original elation R. In vertical partitioning, each site contains only a subset of the attributes of a relation R The application of the classical knowledge discovery process in distributed environment requires the collection of distributed data in a data warehouse for central processing . The following issues are arises while data is distributed [1].

1) Connectivity. Transmitting large quantities of data to a central site may be infeasible.
2) Heterogeneity of sources. Is it easier to combine results than combine sources?
3) Privacy of sources. Organizations may be willing to share data mining results, but not data.

This paper presents secure communication in data mining and distributed data mining. In privacy preserving data mining, data can be protected from third party users using two way enhanced secure approach. This approach includes 1) Perturbation of sensitive

knowledge of data by partial derivatives of functional values. 2) Secure computation of key by the Eigen value of Jacobian matrix which satisfies the implicit function theorem. In secure Two-party computation, There are two parties mutually unknown and they jointly want to compute their data without revealing the data of each other. In Privacy preserving distributed data mining , proposed new secure multiparty communication approach. The basic idea of Secure Multiparty Communication is that a computation is secure if at the end of the computation, no party knows anything except its own input and the results. One way to view this is to imagine a trusted third party { everyone gives their input to the trusted party, who performs the computation and sends the results to the participants. Now imagine that we can achieve the same result without having a trusted party. Obviously, some communication between the parties is required for any **interesting** computation { how do we ensure that this communication doesn't disclose anything? The answer is to allow non-determinism in the exact values sent in the intermediate communication (e.g., encrypt with a randomly chosen key), and demonstrate that a party with just its own input and the result can generate a \predicted" intermediate computation that is as likely as the actual values.[4].

II. NEED OF SECURE MULTIPARTY COMMUNICATION

Secure Multiparty communication (SMC) is a mechanism to provide collaborate computations of multiple organizations without revealing data of individual organization. Each Organization knows nothing except final result. Consider Hospitals, Hospitals generate huge amount of patient data, data must be secured and protect privacy of patient data as per HIPAA rules. Now hospitals would provide better service to the patients, if everyone use the encrypted data and also compute some required function of the data. Based on the given example, consider concept of basic concept of cryptography, in this, there are two persons A and B wants to share their information in secured way. Suppose in this communication, if third person C knows every conversation between them, then how these two persons A and B communicate securely. If C knows everything, then this communication is impossible. So Now A and B have two options for secure communication. a) Share secret data with some key b)Use a secret(Cipher) algorithm. In Multiparty communication, a set of parties with private inputs wish to compute some joint function of their inputs. For example, P: a1 , Q: a2 , R: a3,Supply to a function f(a1 , a2 , a3) which outputs a 3-tuple: (s1 (a1 , a2 , a3), s2 (a1 , a2 , a3), s3 (a1 , a2 , a3)). Sometimes, s1 = s2 = s3, But this is not necessary. In Secure Multiparty communication, a set of parties with private inputs wish to compute some joint function of their inputs. Parties wish to preserve some security properties like privacy and correctness. Security must be preserved in the face of adversarial behavior by some of the participants, or by an external party. The best examples of Secure multi communication are a) Secure election (EVM) b) Auction bidding c) Private Information Retrieval. The Following kind of parties are involved in SMC, such as Trusted third party, Partially trusted parties, Untrusted parties, Competitors. The aim of Secure multi-party computation is allow different parties to carry out their tasks in secure manner[8].SMC provides a base for end-to-end secure multi-party development protocol[7].There are two important properties of SMC; *Privacy and Correctness*. Privacy state that nothing should be learned beyond what is exactly required. Correctness state that each party should receive correct result.

A. Example-Secure Computations with Honest third Party

In ideal Model, group of parties share their original inputs among other parties and sent together with honest third party. When the trusted party returns the output, each majority player outputs it locally, whereas some parties collude compute outputs based on all they know (i.e., the output and all the local inputs of these parties).

A secure multi-party computation with honest majority is required to follow this ideal model. That is, the effect of any feasible adversary which controls a minority of the players in the actual protocol can be essentially simulated by a (different) feasible adversary which controls the corresponding players in the ideal model. This means that in a secure protocol the effect of each minority group is essentially restricted" to replacing its own local inputs (independently of the local inputs of the majority players) before the protocol starts, and replacing its own local outputs (depending only on its local inputs and outputs) after the protocol terminates

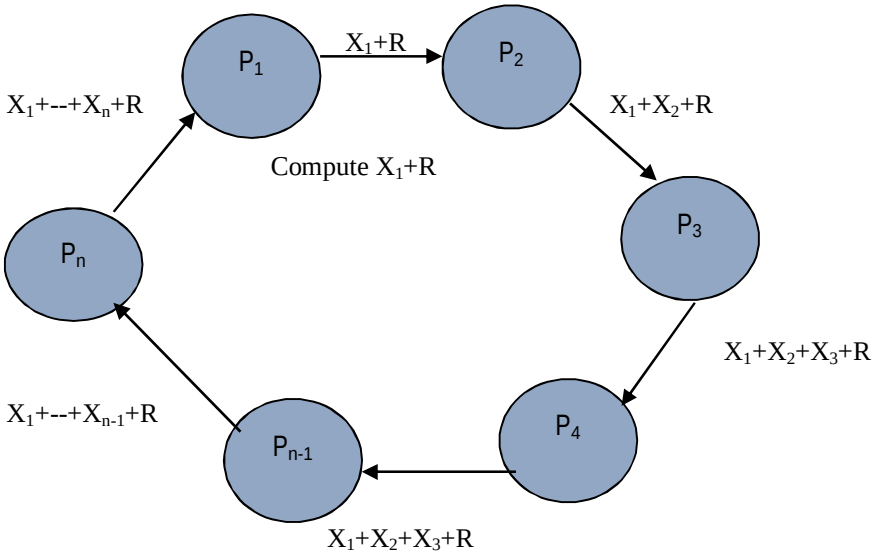
III. SMC TECHNIQUES

A. Secure Sum method using Randomization process to achieve SMC

In Secure Sum method, it is secure if it emulates an idea setting where parties are hand their inputs to a trusted party, who locally computes desired outputs and hand them back to the parties. In the proposed protocol participating parties were organized in a one-way ring. One party works as originator of the protocol through which computation begins by deciding a random number and adding it to its private data. The sum is forwarded to next party for further computation and so on.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

X= Value of P
R=Random Number



Several organizations P_1,----,P_n, wish to perform a combined operation, according to SMC, parties computation should be carried out in this process no organizations can know the input from other organizations. SMC is a technique for PPDM or PPDDM, In which parties collaborate perform a joint computation and each parties only gets the final desired result without knowing the inputs from other organizations. So each organization knows exactly the computation results.

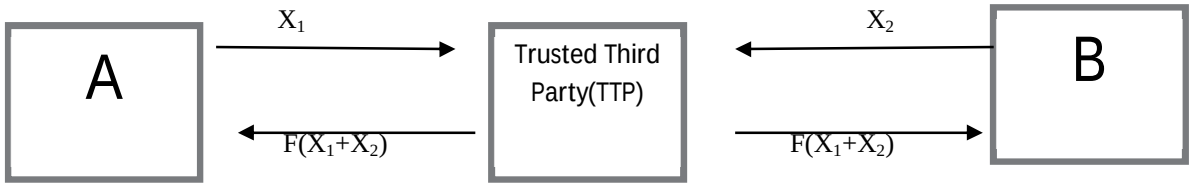


Figure: Ideal Prototype Model of SMC

B. Computation with Randomization and Anonymization

Based on Ideal model of SMC, in this proposed protocol/Method, if parties wish to joint their computation with other parties only through generating pseudo-randomization number(r) on which all the parties mutually agree to performed exactly once. Pseudo-randomization number(r) is used pseudo-randomization function to prevent parties' private data from unauthorized access.

pseudo-randomization function(o,s_i,r) would be divided by one randomly chosen party and distributed among all the other parties and TTP. Once it is received by all the parties, parties will use pseudo-randomization function to hide the actual input then break it in fixed number of packets and distributed among various anonymizers. Anonymizers then forwarded the packet to TTP.TTP recollects the data with the help of pseudo-randomization function to perform the joint computation.

r- pseudo-randomization number(r)

Pseudo-randomization function(o,s_i,r)

Input: (P_1 P_2…..P_n) are parties with (s_1, s_2,s_3 …..s_n) inputs

Output: all the parties and TTP learn f(s_1 s_2....s_n)

1) All the parties jointly executes pseudo-randomization algorithm to generate pseudo-randomization number(r), as a result all the parties learn r.

2) One randomly selected party generates pseudo-randomization function and share it with all the parties and TTP.

3) All the parties then computes

$R(s_1)$=random(o, s_1,r)

$R(s_2)$=random(o,s_2,r)

$R(s_3)$=random(o, s_3, r)

$R(s_n)$=random(o,s_4, r)

4) Each party divided the pseudo-randomized data into fixed number of packets and forward it to randomly selected anonymizer.

5) Anonymizer then forwarded it to TTP.

6) After receiving complete data of all the parties TTP recollects the data and executes joint computation function f(D) and share the results among all the parties.

C. Sharemind Framework

Sharemind [14] is a distributed virtual machine for performing privacy-preserving computations. The Sharemind framework can perform various operations on secret shared 32-bit integers, vectors of 32-bit integers and booleans. The framework allows the developer to write algorithms where public and private data are separated. The Sharemind virtual machine guarantees that private data is not leaked while such an algorithm is evaluated [14].

The Sharemind system uses three servers to hold the shares of secret values. In Sharemind terminology, these servers are data miners. The miners are connected with each other over the network using secure channels and use secure protocols to evaluate a function on the secret shared data. The Sharemind computation protocols are provably secure in the honest-but-curious model with no more than one corrupted party. The honest-but-curious model means that security is preserved when a malicious miner attempts to use the values it sees to deduce the secret input values of all the parties without deviating from the protocol. Secret sharing of private data is performed at the source and each share is sent to a different miner over a secure channel. This guarantees that no-one except the data owner will know the original value [14].

SHAREMIND as a virtual processor that provides secure storage for shared inputs and performs privacy-preserving operations on them. Each miner node P*i* has a local *database* for persistent storage and a local *stack* for storing intermediate results. All values in the database and stack are shared among all miners $P_1, \ldots, P_n$ by using an *additive secret sharing* over $Z_2$32(Sharemind uses additive secret sharing scheme in the ring $Z_2$32 as this allows it to support the efficient 32-bit integer data type) . The current version of SHAREMIND framework is based on three miner nodes and tolerates semi-honest corruption of a single node, i.e., no information is leaked unless
two miner nodes collaborate.

In share computing model, all computational instructions of Sharemind framework are either unary or binary operations over unsigned integers represented as elements of $Z_2$32 or their vectorised counterparts. Hence, all protocols have the following structure. Each miner P*i* uses shares u_i and v_j as inputs to the protocol to obtain a new share w_i such that [[w]] is a valid sharing of $f(u)$ or $u \times v$. In the corresponding idealized implementation, all miners send their input shares to the trusted third party T who restores all inputs, computes the corresponding output w and sends back newly computed shares [[w]] ← Deal(w).Hence, the output shares [[w]] are independent of input shares and thus no information is leaked about the input shares if we publish all output shares

D. Cryptographic SMC Method

Classical approach to SMC is to perform computation using Trusted Third Party (TTP). However, in practical scenario, TTP are hard to achieve and it is necessary to eliminate TTP in SMC. The proposed Elliptic Curve Cryptography (ECC) based approach for SMC that is scalable in terms of computational and communication cost and avoids TTP[15].

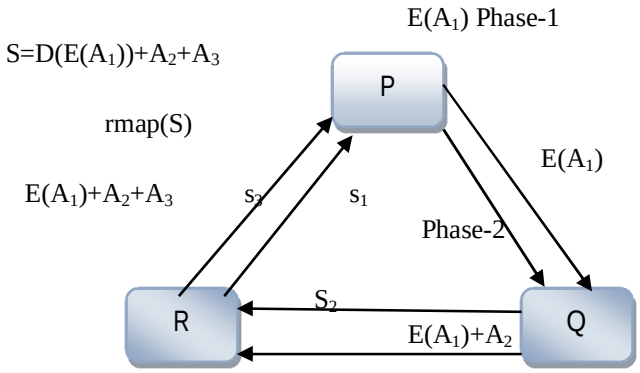
To achieve PPDM, Randomization based and Cryptography based approaches are used. The Cryptography based approaches provide a higher level of privacy but poor scalability [16]. Secure Multiparty Computation (SMC)[17] is a Cryptography-based approach which can be achieved with the help of Oblivious Transfer, Homomorphic and Secret Sharing based schemes. Oblivious transfer based schemes are not scalable due to their high computational and communicational overhead. Secret sharing schemes either use a dedicated server or Trusted Third Party (TTP) to achieve high level of privacy and accuracy but at high computational and communication overhead[18, 19].

The Cryptography based approach provides a higher level of privacy and it can be achieved by the Secure Multiparty Computation (SMC) through Elliptic curve cryptosystem.

In This Proposed approach, We consider the following details: Parties: P,Q,R; private messages: a,b,c; converts private messages into coordinates form A_1, A2,A_3. Encryption/Decryption is performed by Party P. The decrypted total value is sent to Q then to R and P.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

.



A_1 is encrypted by one of the algorithm of ECC which converts the message A_1 to $E(A_1)$ which is sent to party Q. Party Q adds message A_2 with $E(A_2)$ and is sent to R. The cumulative Encrypted with addition of message from R is sent to P. The decryption is then performed on P where we decrypted our encrypted message and we get the total message as S. We performed rmap(S) by which we get the message s_1(The general frame- work for SMC consists of specifying a random process that maps m inputs (local inputs of parties) to m outputs (desired outputs)). This message from P is given to Q and subsequently to R and then back to P.

The above process can be divided into two phases. In Phase-1 the initiator first encrypts its private value using ECC based encryption scheme. The resultant cipher text (which is in the form of elliptic curve point) is sent to next party in the ring. Next party does not perform any cipher operation but just adds its own private value (mapped to elliptic curve point) with the received cipher text. This process is repeated and finally initiator receives the message $E(A_1) + A_2 + A_3$ at the end of phase I.

In phase II, initiator decrypts the message by removing the noise (that was added during encryption) from the message and computes A1+ A2 + A3. Here, initiator just removes the noise in order to get desired sum

IV. RELATED WORK ALGORITHMS

A. Two way method enhanced security approach

1) The first one is to changing original data prior to passing through to the data mining system. so that real values of data are ambiguous(perturbation).

2) Privacy preserving Data mining by using Implicit Function Theorem is two way method enhanced security approach includes:

a) Data modification using Perturbation of sensitive knowledge of data by partial derivatives of functional values.

b) Secure computation of key by the Eigen value of Jacobian matrix which satisfies the implicit function theorem

B. Algorithm: Privacy Preserving Data Mining By Using perturbation method

1) *Input:* Extracted data mining result or input data base to be shared

2) *Output:* Generate dynamic secret key to encrypt and decrypt the data.

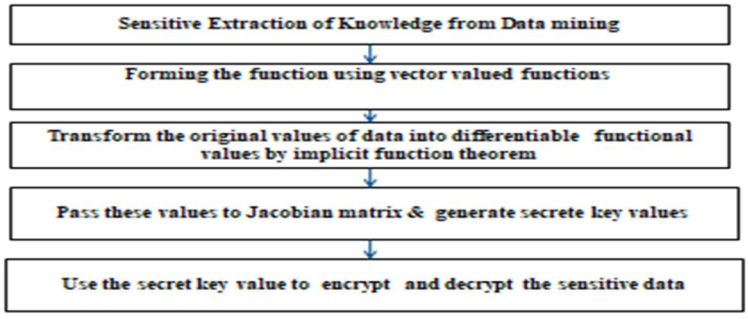
a) *Step 1:* Identify the sensitive values from the input data

b) *Step 2:* Form the sensitive data values into vector valued functions

c) *Step 3:* Achieve perturbation by transforming the sensitive values of data into differentiable functional values by Implicit function theorem.

d) *Step 4:* Pass these values to Jacobian matrix and generate Eigen values

e) *Step 5:* Select random Eigen secret key value to encrypt and decrypt the sensitive data.



C. The second approach is privacy preserving distributed data mining(secure multi-party computation).

1) Privacy preserving distributed data mining(secure computation): Privacy preserving distributed K-means clustering using elliptic curve cryptography(ECC).

2) K-Means clustering aims to partition n records into k clusters in which each record belongs to the cluster with the nearest center.

3) Distributed K-Means clustering requires computation of (sum of records) /(number of records) as an intermediate step to compute global cluster means in each iteration.

4) The algorithm proceeds by alternating between two steps:

5) Assignment step: For each record, calculate the distance of a record to all k means and assign record to the cluster with the closest mean.

6) Update step: Calculate the new means to be the centroid of the records in the cluster.

D. Algorithm: Privacy preserving distributed K-means clustering using elliptic curve cryptography.

1) Step1: Partition n records into k clusters

2) Step2: computation of (sum of records)/(number of records) to compute global cluster means in each iteration.

3) Step3: Method works in two phases;

4) Step 4: Phase-1, each party performs local clustering and computes sum of records and number of records values for each cluster.

5) Step 5: Phase-2, The initiator Upon receiving message, initiator removes the noise added to the message during Phase I and gets original data.

6) Step6: Initiator forwards this sum to the next party in the ring and subsequently all parties receive the sum. As the sum is a point on the curve, all parties remap this sum and get the original sum.

7) Step7: Step6 repeated for all sum of records and number of records values and finally at the end of the iteration, all parties will be able to compute (sum of records)/(number of records) and hence global cluster means.

V. CONCLUSIONS

In this paper we describe Secure Multi-party computation and its techniques. In techniques part describes about different approaches to achieve sharing computations without revealing the original and shared data. This paper also includes the importance of SMC for privacy preserving computations and algorithms. Future scope of this paper is developing these techniques in practical mode and can use it for any real world task analysis.

REFERENCES

[1] Chris Clifton, Privacy Preserving Distributed Data Mining.
[2] V.THAVAVEL and S.SIVAKUMAR, A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment, IJCSI, Vol. 9, Issue 1, No 2, January 2012.
[3] V. Baby, N. Subhash Chandra, Privacy-Preserving Distributed Data Mining Techniques:A Survey, International Journal of Computer Applications (0975 – 8887)Volume 143 – No.10, June 2016.
[4] Chris Clifton, Murat Kantarcioglu,Jaideep Vaidya,Xiaodong Lin, Michael Y. Zhu, Tools for Privacy Preserving Distributed Data Mining,SIGKDD Explorations, Volume 4, Issue 2
[5] Yehuda Lindell and Benny Pinkas, Secure Multiparty Computation for Privacy-Preserving Data Mining, The Journal of Privacy and Confidentiality (2009).
[6] R. Sugumar , Dr. C. Jayakumar and A. Rengarajan, Design a Secure Multiparty Computation System for Privacy Preserving Data Mining,IJCST, Volume 3, Issue 1, January 2012.
[7] Walter Priesnitz Filho and Carlos Ribeiro, State of the art of secure multiparty computation for privacy preserving data mining, Geintec - ISSN: 2237-0722. Aracaju/SE. Vol. 7, n.4, p.4131-4148, out/nov/dez – 2017.
[8] Anand R. Padwalkar, Secure multiparty computation protocol: Basic building blocks methods, IJCES, Volume 4, Issue 2,March-2014.
[9] Du, Wenliang and Atallah, Mikhail J., "Secure Multi-Party Computation Problems and Their Applications: A Review And Open Problems" (2001). Electrical Engineering and Computer Science. Paper 11. http://surface.syr.edu/eecs/11.
[10] U. Kumaran and Neelu Khare, A Review on Privacy Preserving Data Mining using Secure Multiparty Computation, Indian Journal of Science and Technology, Vol 9(48), December 2016.
[11] Jessie Covington and Megan Golbek, SECURE MULTIPARTY COMPUTATION, REU program in Mathematics at Oregon State University supported by NSF grant DMS-1359173, August 2015.
[12] Prof. Anand R.Padwalkar, Secure Multiparty Computation Protocol used for Privacy Preserving Data Mining-Zero Data Leakage Approach, International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 3 Issue 1 September 2013.

[13] Hirofumi Miyajima, A proposal of privacy preserving reinforcement learning for secure multiparty computation, Artificial Intelligence Research, 2017, Vol. 6, No. 2.
[14] Dan Bogdanov, Sven Laur, and Jan Willemson. Sharemind: A Framework for Fast Privacy-Preserving Computations. In Sushil Jajodia and Javier Lopez, editors,Computer Security { ESORICS 2008, volume 5283 of Lecture Notes in Computer Science, pages 192{206. Springer Berlin / Heidelberg, 2008.
[15] Ankit Chouhan, Sankita Patel, Dr. D. C. Jinwala, Comparative Analysis of Elliptic Curve Cryptography Based Algorithms to Implement Privacy Preserving Clustering through Secure Multiparty Computation, published in Journal of Information Security, Scientific Research, 2013.
[16] Wang, Liu, Yue. 2007 Privacy preserving data mining re-search: current status and key issues. In: 7th International Conference on Computational Science 2007, pp. 762–772.
[17] O.Goldreich. 2004. The Foundations of Cryptography, vol. 2. Cambridge Univ. Press, Cam-bridge.
[18] Upmanyu, M., Namboodiri, A.M., Srinathan, K., Jawahar. 2010. Efficient Privacy Preserving K-Means Clustering. In: PAISI, pp.154-166.
[19] Doganay, Pedersen, Saygin, Savas. 2008. Distributed privacy pre-serving k-means clustering with additive secret sharing. In: 2008 international workshop on Privacy and anonymity in information society, pp. 3-11. Nantes, France.
[20] http://www-ai.cs.uni-dortmund.de/auto?self=$ejr31cyc