Real Time Host Assessment using ANN for Grid Network

Prof. T.R.Srinivasan
Department of IT
Vidyaa Vikas College of Engineering and Technology
Tiruchengode. India.

Dr.R.Shanmugalakshmi
Department of CSE & IT
Government College of Technology
Coimbatore, India.

***Abstract*— In this work, we study the efficacy of assessing registered hosts for job allocation using Artificial neural network to classify registered hosts during job scheduling.**

Grid is evolving as the computing structure of the future. The success in commercial grid computing is the ability to negotiate resource sharing arrangements with a set of registered participating parties. Grid computing is capable of integrating services across distributed heterogeneous disparate resources with a centralized control to provide quality of service.

Host assessment plays a crucial role to assign a specific job in the grid. Host selection among the registered pool of hosts can drastically improve the quality of service. Resource discovery algorithms are available but identifying ideal resource to reduce queue time and response time is the most essential task in a commercial grid environment.

Resource mining is the process of running data through sophisticated algorithms to uncover meaningful patterns and co relations that may otherwise be hidden. We explore the application of these techniques to assess host by training the system with known data.

Experimental results show an improvement of 25.49 percent in data classification using ANN over normal methods.

Keywords— ***Grid computing, Artificial Neural Network, Distributed computing, host assessment, Weka***

I. Introduction

The availability of Internet, the availability of higher end computers along with very high speed network has made a huge difference in the minds of the people towards the use of the computers in their day-to-day life. Now people are thinking of finishing their jobs at a much faster rate. To accomplish the needs of the each and every individual in terms of computational resource, it is very hard to have everything within a compound. A solution for such a problem can be a Grid computing network [1]. Grid is a dynamic network which is geographically distributed throughout the world and not owned by an individual organization. The dynamic nature itself is the advantageous concept in it because the users can use any high rated resources which are available during a particular period of time.

Grid network is a computational network which can be dynamically constructed based on factors like interest of the client user, decision taken by the server based on certain decision algorithm, processing speed, transmission bandwidth and memory available in the client machine and so on. We cannot predict how the grid network will be since it is going to be shared by many client users. Here the client may enter or exit the network at any time depending upon his requirement[2][3].

A grid server typically has modules including Grid Scheduler, Job Assigning & Monitoring, and Grid Information Service. Figure 1 shows the Grid scheduling architecture .

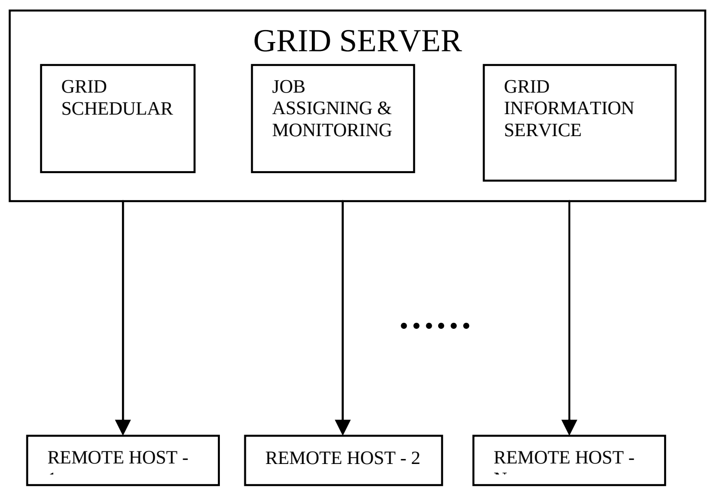

Figure 1: Grid Scheduling Architecture

A typical grid network has its own scheduler for the effective use of its client resources satisfying the associated performance and policy constraints which will be assigned while designing the grid network. The Job assigning & monitoring module will take care of assigning and monitoring jobs that has been allocated for various local applications in the remote machine. It is also responsible for reallocation of jobs in case of requirements [4]. Grid information service will provide the necessary information and statistical data about the various clients those who had registered with this grid network. Local applications are those applications which will be connected to the grid server through the Internet. Some local applications will request the server for services while some others might be providing service to the server[5][6][7].

II. DATA MINING

With the continuous development of database technology and the extensive applications of database management system, the data volume stored in database increases rapidly and in the large amounts of data much important information is hidden. If the information can be extracted from the database they will create a lot of potential profit for the companies, and the technology of mining information from the massive database is known as data mining.

Data mining tools can forecast the future trends and activities to support the decision of people. For example, through analysing the whole database system of the company the data mining tools can answer the problems such as “Which customer is most likely to respond to the e-mail marketing activities of our company, why”, and other similar problems. Some data mining tools can also resolve some traditional problems which consumed much time, this is because that they can rapidly browse the entire database and find some useful information experts unnoticed.

The application of neural networks in the data mining [8][9] is becoming popular. Although neural networks may have complex structure, long training time, and uneasily understandable representation of results, neural networks have high acceptance ability for noisy data and due to its high accuracy is preferable in data mining.

Neural Networks [Starsoft, inc] are analytic techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called *learning* from existing data.

The first step is to design a specific network architecture (that includes a specific number of "layers" each consisting of a certain number of "neurons"). The size and structure of the network needs to match the nature of the investigated phenomenon. Because the latter is obviously not known very well at this early stage, this task is not easy and often involves multiple "trials and errors." The new network is then subjected to the process of "training." In that phase, neurons apply an iterative process to the number of inputs (variables) to adjust the weights of the network in order to optimally predict the sample data on which the "training" is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions. The resulting *"network"* developed in the process of *"learning"* represents a pattern detected in the data. Thus, in this approach, the *"network"* is the functional equivalent of a model of relations between variables in the traditional *model building* approach. However, unlike in the traditional *models*, in the *"network,"* those relations cannot be articulated in the usual terms used in statistics or methodology to describe relations between variables (such as, for example, *"A is positively correlated with B but only for observations where the value of C is low and D is high"*). Some *neural networks* can produce highly accurate predictions; they represent, however, a typical a-theoretical (one can say, "a black box") research approach. That approach is concerned only with practical considerations, that is, with the predictive validity of the solution and its applied relevance and not with the nature of the underlying mechanism or its relevance for any "theory" of the underlying phenomena. Figure 2 shows the neural network model.

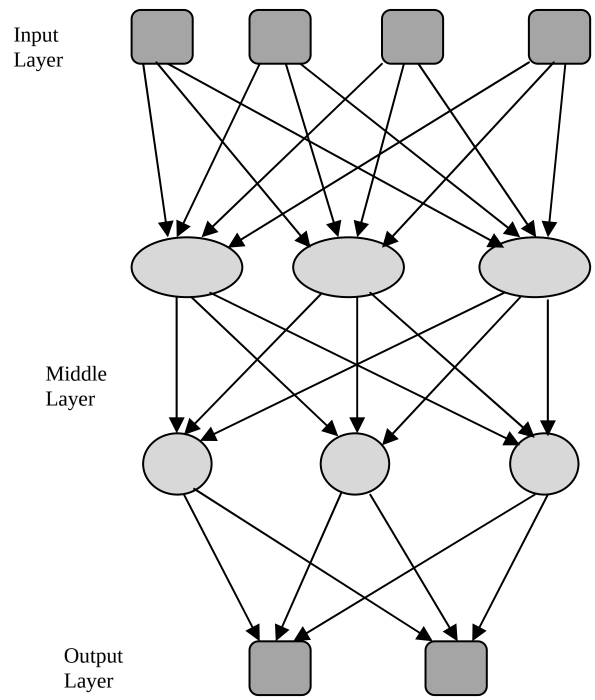

Figure 2. A neural network model

However, it should be mentioned that *Neural Network* techniques can also be used as a component of analyses designed to build explanatory models because *Neural Networks* can help explore data sets in search for relevant variables or groups of variables; the results of such explorations can then facilitate the process of model building.

One of the major advantages of *neural networks* is that, theoretically, they are capable of approximating any continuous function, and thus the researcher does not need to have any hypotheses about the underlying model, or even to some extent, which variables matter. An important disadvantage, however, is that the final solution depends on the initial conditions of the network, and, as stated before, it is virtually impossible to "interpret" the solution in traditional, analytic terms, such as those used to build theories that explain phenomena.

III. DATASET USED IN OUR RESEARCH WORK

In this paper we investigate ten attributes containing features of the nodes in the grid including available memory, cache, processor speed and available bandwidth. 300 node attributes were used of which 66% was assigned to training set and the remaining for validation. the available bandwidth and memory were created using a random seed. the class label grid was given two values compute and data. compute hosts at a given instance will provide access to huge pool of shared processing power suitable for high throughput applications and computation intensive computing. data host at a given instance provide an infrastructure to support data storage, data discovery, data handling, data publication, and data manipulation of large volumes of data actually stored in various heterogeneous databases and file systems.

IV.ANALYSIS OF KEY ATTRIBUTES

The statistical description of the processor and the main board of the nodes used in our work is shown in table 1.

	processor	speed	cache	FSB
Median	2400.00	512.00	533.00	512.00
Mode	2800	1024	400	512
Std. Deviation	544.156	402.963	168.287	536.130
Skewness	-.267	1.207	.459	1.235

Table 1. Statistical description of the nodes used in our work.

The frequency distribution and cumulative percentage of some of the discovered node attributes is represented in table2, 3,4 & 5.

Processor speed	Frequency	Percent	Cumulative Percent
256	53	17.7	17.7
512	116	38.7	56.3
1024	119	39.7	96.0
2048	12	4.0	100.0
Total	300	100.0	

Table2 : Frequency distribution of processor speed

processor speed	Frequency	Percent	Cumulative Percent
1400	1	.3	.3
1600	43	14.3	14.7
1800	49	16.3	31.0
2000	6	2.0	33.0
2400	59	19.7	52.7
2600	7	2.3	55.0
2800	73	24.3	79.3
3000	39	13.0	92.3
3200	22	7.3	99.7
3600	1	.3	100.0
Total	300	100.0	

Table 3. Frequency distribution of Cache

Front Side Bus	Frequency	Percent	Cumulative Percent
256	2	.7	.7
512	153	51.0	51.7
1024	99	33.0	84.7
2048	46	15.3	100.0
Total	300	100.0	

Table 4. Frequency distribution of processor speed

Cache	Frequency	Percent	Cumulative Percent
400	113	37.7	37.7
533	91	68.0	30.3
800	96	100.0	32.0
Total	300		100.0

Table 5. Frequency distribution of Front Side Bus

IV. CLASSIFICATION MODEL FOR REMOTE HOST CLASSIFICATION IN A GRID NETWORK

We use the artificial neural network classifier and compare the results with Clonalg algorithm. The flow chart of the assessment is given in figure 3.

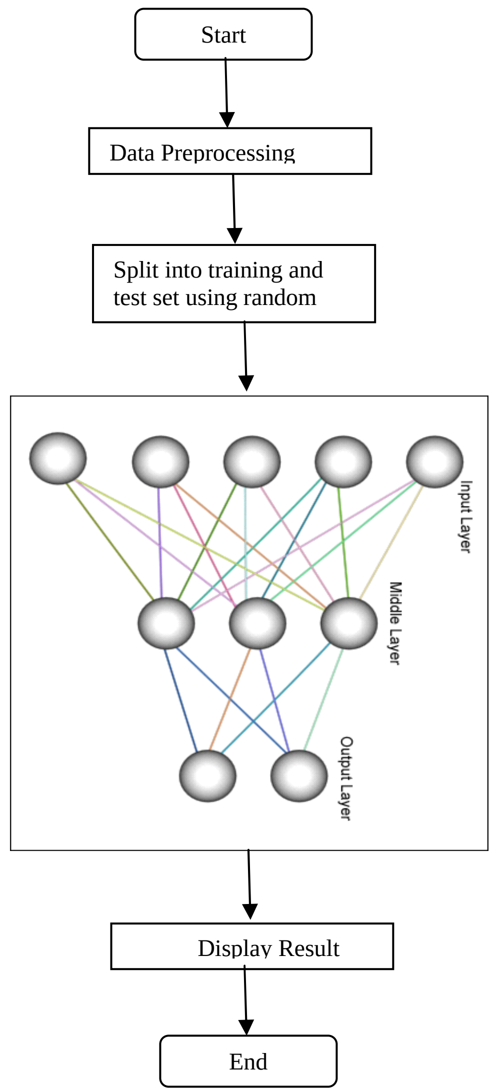

Figure 3. Flow chart of the mining process.

V. RESULT AND CONCLUSIONS

WEKA was used to analyse our data. We compare two classification methods namely CLONALG and Neural network. From figure no 4, we can see the classification obtained for different strategies.

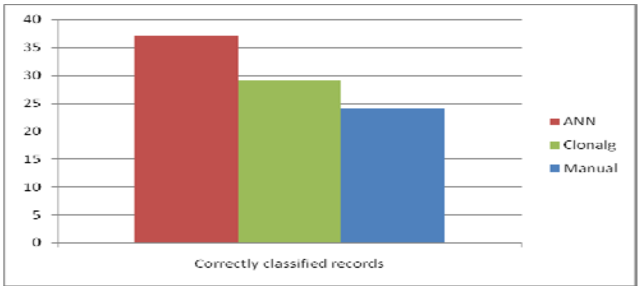

Figure 4. Classification accuracy of various methods

We have been able to achieve good classification results using ANN algorithm (72.55%) which shows that resource selection can be drastically improved over manual (47.05%) or CLONALG (56.86%) methods. Though AIS is an emerging field its application in Grid computing needs to be analysed further.

REFERENCES

[1] Fangpeng Dong and Selim G. Akl, Technical Report No.2006-504 Scheduling Algorithms for Grid Computing: State of the art and open problems, School of Computing, Queen's University, Kingston, Ontario, January 2006.

[2] Liang Peng; See, S.; Yueqin Jiang; Jie Song; Stoelwinder, A.; Hoon Kang Neo;Performance evaluation in computational grid environments,

[3] High Performance Computing and Grid in Asia Pacific Region, 2004. Proceedings. Seventh International Conference, Publication Date: 20-22 July 2004.

[4] Afzal, A.; Darlington, J.; Mcgough, A., QoS-Constrained Stochastic Workflow Scheduling in Enterprise and Scientific Grids; Grid Computing, 7th IEEE/ACM International Conference on 28-29 Sept. 2006..

[5] Xiaowu Chen; Mingji Chen; Xixi Luo , A Multilayer Portal Model for Information Grid,; ChinaGrid Annual Conference, 2008. ChinaGrid '08. The Third, 20-22 Aug. 2008 Page(s):78 – 85.

[6] Liang Peng; Koh, M.; Jie Song; See, S , Grid Service Monitoring for Grid Market Framework, Networks, 2006. ICON '06. 14th IEEE International Conference on Volume 1, Sept. 2006 Page(s):1 – 6

[7]Do-Hyeon Kim; Kyung-Woo Kang, Design and Implementation of Integrated Information System for Monitoring Resources in Grid Computing; Computer Supported Cooperative Work in Design, 2006. CSCWD '06. 10th International Conference on 3-5 May 2006 Page(s):1 – 6

[8] Xianjun Nihas, Research of Data Mining Based on Neural Networks,

[9] Jason Brownlee, Artificial immune recognition system: a review and analysis, The clonal selection classification algorithm, Technical Report 2-02