2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

A Multi-Agents Approach to Knowledge Discovery

Cuong Tong, Dharmendra Sharma and Fariba Shadabi
Faculty of Information Sciences and Engineering,
University of Canberra, ACT, 2601Australia

Abstract

Over the past few years, data mining and multi-agent approach has been used successfully in the development of large complex systems. Such a hybrid approach can be considered as an effective approach for the development of predictive modeling in complex e-health systems. We propose a real time Data Mining and Multi-Agent System called DMMAS, modeling chronic disease data. DMMAS approach employs data partitioning and multiple agents with option to employ heterogeneous or homogenous data mining techniques, distributing agent based processing for modeling and combining results from all the agents to improve the efficiency.

1. Introduction

Multi-agent and data mining have been widely used in building large and complex system. A group of agents can collectively and collaboratively form a Multi Agent System (MAS) to perform complex and lengthy tasks [1].

As a multidisciplinary field, data mining draws from areas such as artificial intelligence, database theory, data visualization, marketing, mathematics, operations research, pattern recognition, and statistics. Research into data mining has thus far focused on developing new algorithms and on identifying future application areas [2].

Han and Kamber [2] identify one of many problems in data mining is performance issues which include efficiency, scalability and parallelization of data mining algorithms. They suggest that to effectively extract information from a huge amount of data in database, data mining algorithm must be efficient and scalable. In another word, running time of a data mining algorithm must be predictable and acceptable in large database.

The computational complexity of some data mining methods can be considered as the most important factors that motivated the development of parallel and distributed data mining algorithms. In parallel, distributed and incremental mining algorithms, data can be divided into partitions that are processed in parallel. The results from the partitions are then merged.

This paper proposes a new approach to address some of the above limitations and concerns, with strong focus on efficiency. The system can assist in data mining on common chronic disease data. In the context of this paper, type II diabetes is selected as the study domain because of growing number of people being affected each year [3].

The basic hypothesis for this approach is to explore if we can employ multiagent system to improve large and complex data mining task.

2. Background

It was shown that chronic disease is reaching epidemic proportions in developed nations around the world. The number of people with chronic illness is expected to grow rapidly over the coming decades as population age and adverse trends in lifestyle behaviors persist into the future [4]. According to [3, 4] if current trends in population ageing and modifiable risk factors for type 2 diabetes continue over the next 20 years in Australia, then the number of Australians aged 25 years and over who will have type 2 diabetes will increase by over 70% from around 1 million to 1.8 million people. The best way to improve diabetes patients' longevity and reducing complication is to detect it as early as possible.

Diabetes is chronic health condition where the body is unable to automatically regulate blood glucose levels, resulting in continual elevated glucose level in the blood. "Insulin" is a chemical hormone in pancreas. If diabetes is left undiagnosed or poorly treated it increases the chances of complications, which include heart disease, kidney disease, nerve and circulation damage, impotence, blindness and lower limb amputation [3].

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.418

Data mining in general falls in to the following categories: classification patterns, association patterns, sequential patterns, and spatial-temporal patterns. In diabetic database application, the focus is on the mining of association patterns. Classification patterns provide a description of the characteristics of the population having certain diabetic attribute. Association patterns provide a list of symptoms or treatments that often occur together. [5]

There are researches devoted to data mining in diabetes such as [6], p 430 – 436, [5] carry out knowledge discovery on large diabetic patient database and presented some lessons learnt.

One of the challenges with data mining is the performance. Performing data mining on a large dataset can take very long time. In addition, the data analysis is not up to date with the constant incoming data which result in invalid result being made. There are also other studies focusing on data mining performances on large data set [7], [8]. Other similar approach such as CoLe a cooperative data mining approach for discovering hybrid knowledge. It employs multiple different data mining algorithms, and combines results from them to enhance the mined knowledge [9].

3. Data mining in real time

Data mining in real time is aligned with Westphal et el.'s definition for data monitoring where it involves the processing of data that are continually being updated [10]. What is true one moment may abruptly become out of date and invalid the next moment. Monitoring often make quick response in order to take advantage of information as it is being presented. As a result, predictive models and for casters can be used to help identify critical values, unusual behaviour and criteria data [10].

Typically, the dataset D will consist of some initial data, N rows. The number of rows is divided into I partitions. The number of partition can be varied between 1 to N (I = N/q where q = 1 to N).

Base on the number of data partitions, the same numbers of agents are created to take ownership of that data partition. This agent is then an expert and responsible its data partition (eg DataMiningAgent n (DM_n) is responsible for data partition n_{th})

The number of partitions to be generated can be specified in DMMAS.

When the very first time DMMAS runs, it is necessary to specify the data set to work with. A number of partitions will be generated that assembles the original dataset.

Each Data Mining Agent performs its own data mining task and generates its own rule set bases on the data partition that it manages.

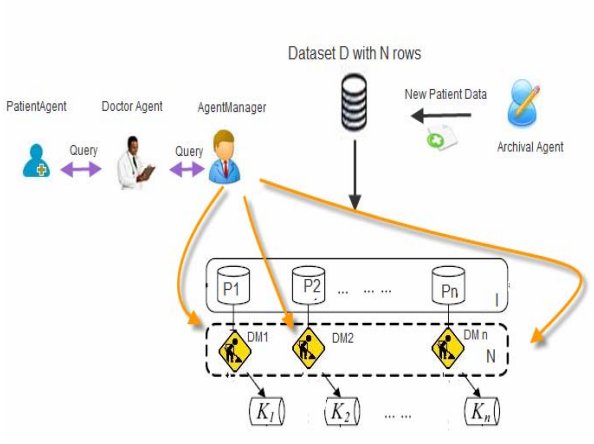

Figure 1: Conceptual view of DMMAS

The following sequence illustrates the activities involved when patient agent sends a query to DMMAS

1. The patient agent sends a query that contains patient's attribute to DoctorAgent
2. DoctorAgent seeks advice from AgentManager
3. AgentManager broadcasts the query to all active data mining agents for consultation.

a. Data mining agents return the result base on their local knowledge.

4. The AgentManager consolidates data from all agents and presents the answer back to the DoctorAgent.
5. DoctorAgent sends the result back to the PatientAgent.

As a certain number of new data is added to the dataset, new agent is created and is allocated a dataset. The data mining efficiency will improve because of its distributed architecture. In addition, as new data is being added; existing models do not always have to be regenerated. The challenge here is to find out the optimal number of partition (q) where we can still maintain a reasonable level of accuracy and efficiency. This can be achieved with a trial and error where a fine tuning is carried out by experiencing different value of partition. At the current stage, data mining agents are restricted to homogenous algorithms for data mining tasks.

4. Why multi agents system

Multi agents system is selected for the proposed system for several reasons.

First of all, data mining is a large complex task, highly dynamic and uncertain. It is large in scale and can be distributed spatially. Multi agents can be used to solve problems that are too large for a centralized agent to solve because of resource limitations and/or to avoid a one point bottleneck or failure point [11], [12] [1].

Second, agents are capable of independent actions on behalf of a user or owner and can act, capture and manage information automatically when it is necessary.

Thirdly, agents can interact with other external systems and can be used to manage both distributed and local knowledge. This is an important feature since E-health knowledge is usually generated by different sources and often from different places. For instance, a consultation agent can work with diagnosis agent in order to provide a better answer to the enquiry of patients.

Fourthly, agents can learn from their own experience. This is particularly important in the field of data mining as the data is constantly modified and updated. This results in the system to perform better over time since the agents have learnt from their previous experiences.

Finally, agents have the autonomy and social ability, and multi-agent system is inherently multi-threaded for control. Therefore, multi-agent approach is very effective for tackling the complexity of e-medicine systems and therefore is suitable for the development of e-medicine systems. [2]

5. Result

The experiment was conducted using the UCI Pima Indian data set [3]

The dataset contains 768 instances of Pima Indian heritage females who were diagnosed for diabetes. There were 268 instances diagnosed with diabetes. There are 8 attributes and the diagnostic result (diabetes negative or diabetes positive) in the data set.
The attributes are as follow: number of times pregnant, plasma glucose concentration, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), body mass index, diabetes pedigree function, age (years) and finally the test result as class variable (0 or 1)

The experiment is to measure time and accuracy of a classification on the UCI Pima Indian dataset [15]. The target variable is Class Variable (number 9) in the dataset. Class variable value is mutually exclusive, either diabetes negative or diabetes positive.

The classification algorithm used for training is Quinlan's C4.5 decision tree.
As the experiment focuses on the performance, it is essential for the dataset to be sufficiently large to see some meaningful improvement.

Due to the constraint size of available dataset, a new dataset is created from the original dataset (i.e. UCI Pima Indian dataset) where additional training examples was generated by using a random sampling with replacement strategy.

In the experiment, 5 random datasets from UCI Pima Indian are generated with the size of 2000, 10000, 20000, 50000 and 100000 instances. Each dataset is then used in single classifier and multiagent classifiers for training and testing. The results are recorded for comparison.

Single classifier

This is where a single process performs training and testing on the generated dataset. A dataset is divided into two parts: training and testing. Sixty six percent of the dataset is selected randomly for training and the remaining thirty three percent is used for testing.

Multi-agent classifiers

For each generated dataset above, the dataset is divided into 4 partitions, 10 partitions, 20 partitions and 50 partitions respectively. Each agent is allocated a distinct partition. The agent then train classifier and test classifier for that partition. The same algorithm used in single classifier is applied for each agent. The classification duration and accuracy of each agent are averaged and recorded in the following tables.

Experiment infrastructure

The experiment runs on the following infrastructure

Hardware	
Processor	Intel Core 2 Duo 2Ghz
Memory	2 GBs
Hard disk	120 GBs

Software	
Operating System	Microsoft Window Vista Business 32 Bits
Java runtime	1.6.010-beta
Agent Framework	JADE 3.6
Data mining library	Weka 3.5.7
External Library	Agent Academy 2.0
Development Environment	Eclipse 3.3.2

The measured duration excludes data source loading time and training model saving time. The measurement starts when the training process begins and stops when the testing is completed.

The experiment is repeated 5 times. The total duration per run is slightly different to each other. The average time taken and accuracy of each runs are

recorded and averaged. In the multi-agent classifier, each agent's duration and accuracy are sum up and averaged each time. The initial experiment demonstrates that multiagent classifiers can improve the algorithm efficiency. However, this comes with a cost of slightly lower accuracy.

Experiment with 50 agents demonstrated the lowest running time (1610ms) with little loss of accuracy (96.23%) in comparison to the traditional approach (12088ms, 100%); multi-agents approach with 20 agents (2304ms, 98.55%). There is no concrete formula on the number of speed improvement with a certain number of agents. However, base on the experiment result, for C4.5 algorithm, with 4 agents the speed is double and the accuracy loss is 1%. With 50 agents the speed is 7.5 times faster with accuracy loss is 9.8%. Multiagent system classifier can become inefficient and inaccurate if there are a large number of partitions. There is not yet conclusion on what is the best number of agents for a certain dataset size.

6. Conclusion

Research shows that over the past few years there has been great interest in the use of data mining tools across the healthcare spectrum. In this study, we proposed a real time data mining cooperative multi agents system called DMMAS. DMMAS is a multiagent system with multiple miner agents and a combination agent as agent manager. The main goal the system is to explore how data partitioning and multi agent approach can help to improve the efficiency and also if possible the accuracy of chronic diseases management and prediction tasks in real time. Our initial experimental results have shown promising results, in this case, using the diabetes data. It should be noted that the application domain is not limited to diabetes data and it can be applied across other domains for improving the quality of care.

7. Future works

Running DMAAS on better infrastructure, such as 64 bit operating system to leverage 64bit memory addressing can significantly improve DMMAS efficiency. The experiment conducted assumes the data is clean and contains no missing value. In real world this is not always the case. The data population in the experiment is limited. It is desirable to use a large data file for training and testing purposes.

Each agent may utilize different reasoning techniques depending on the situation. For instance, data mining agent 1 can use a decision tree algorithm, data mining agent 2 can use regression, data mining agent 3 can use support vector machine.

There are still many challenges to be overcome. Each technique usually has its own advantages and disadvantages under different circumstances. The comparison study is, however, outside the scope of this paper although it merits further research.

8. References

[1] M. Wooldridge, *An Introduction to MultiAgent Systems. John Wiley & Sons Ltd, 2002.* John Wiley & Sons, 2002.

[2] J. Han and M. Kamber. (2001, *Data Mining : Concepts and Techniques.* Available: http://www.loc.gov/catdir/description/els031/00042822.html, http://www.loc.gov/catdir/toc/els031/00042822.html FLG

[3] IDI. International diabetes institute - diabetes research, education and care. *2007(10/30/2007),*

[4] Laurie J Brown, Anthony Harris, Mark Picton, Linc Thurecht, Mandy Yap, Ann Harding and Peter Dixon, "Linking Microsimulation and Macro-Economic Models to Estimate
the Economic Impact of Chronic Disease Prevention," 2007.

[5] W. Hsu. (2000, Exploration mining in diabetic patients databases: Findings and conclusions.

[6] R. Ramakrishnan and S. Stolfo. (2000, *KDD-2000: Proceedings, August 20-23, 2000, Boston, Massachusetts, USA.* Available: http://isbndb.com/d/book/kdd_2000

[7] J. H. Dula. (2008, A computational study of DEA with massive data sets. *Computers operations research 35(4),* pp. 1191.

[8] P. Buneman and S. Jajodia. (1993, *SIGMOD '93.* Available: http://isbndb.com/d/book/sigmod_93

[9] J. Gao, J. Denzinger and R. C. James. (2005, CoLe: A cooperative data mining approach and its application to early diabetes detection. *Icdm 0*pp. 617-620. Available: http://doi.ieeecomputersociety.org/10.1109/ICDM.2005.44

[10] C. R. Westphal and T. A. Blaxton, *Data Mining Solutions: Methods and Tools for Solving Real-World Problems.* New York, USA: John Wiley & Sons, 1998,

[11] S. R. Ira. (2004, INTELLIGENT AGENTS. *Communications of the Association for Information Systems Volume14, 2004*pp. 275-290.

[12] D. Sharma and F. Shadabi, An Intelligent Multi Agent Design in Healthcare Management System- Lecture Notes in Computer Science, Volume 4953/2008, pp. 674-682, 2008.

[13] N. R. Jennings and M. Wooldridge, "Applications of intelligent agents," in *Agent Technology: Foundations, Applications, and Markets*, N. R. Jennings and M. Wooldridge, Eds., 1998, pp. 3-28.

[14] J. Tian and H. Tianfield, "A Multi-agent Approach to the Design of an E-medicine System," in *Multiagent System Technologies*, 2003, pp. 1093-1094.

[15] U. P. Indian, "Pima Indians Diabetes Data Set." vol. 2008, 2008.