The Adaptability of Conventional Data Mining Algorithms through Intelligent Mobile Agents in Modern Distributed Systems

Dost Muhammad Khan[1], Nawaz Mohamudally[2]

[1]*Assistant Professor, Department of Computer Science & IT, The Islamia University of Bahawalpur, Pakistan & PhD student, School of Innovative Technologies & Engineering, University of Technology, Mauritius*

[2]*Associate Professor, & Consultancy & Technology Transfer Centre, Manager, University of Technology, Mauritius (UTM)*

Abstract

Intelligent mobile agents are today accepted as powerful tools for data mining in a distributed environment. The use of data mining algorithms further beefs up the intelligence in software agents. Knowledge discovery and data mining algorithms are applied to discover hidden patterns and relations in complex datasets using intelligent agents. The distributed computing provides remote communication, fault tolerance, high availability and remote information access. The uses of intelligent mobile agents using data mining algorithms make distributing computing more popular and acceptable. This paper is about the use of conventional data mining algorithms through intelligent mobile agents in distributed systems.

Keywords: *Data Mining, Distributed Systems, Multiagent System (MAS)*

1. Introduction

The ever growing amount of data that are stored in distributed form over networks of heterogeneous and autonomous sources poses several problems such as network bandwidth, communication, autonomy preservation, scalability, data buffering and privacy protection. The techniques of mining information and knowledge from huge data sources such as weather databases, financial databases or emerging diseases information systems were originally developed for handling the distributed cases. The increasing demand to extend data mining technology to datasets inherently distributed among a large number of autonomous and heterogeonous sources over a network with limited bandwidth has motivated the development of new approaches in distributed systems and knowledge discovery. A few of these approaches make use of intelligent mobile agents. Due to the adaptive and deliberative reasoning features of intelligent agents, they are well suited to cope up with the problems of distributed systems. An intelligent, learning and autonomous agent is capable of capturing and applying domain specific knowledge, learning, information and reasoning, to take actions in pursuit of a goal. The data is available every where now. The issue is to produce knowledge from this mountain of data. This is an era of ‘knowledge’ and ‘knowledge’ can be obtained with the combination of ‘data’ and ‘information’ by using data mining techniques. Data mining is successful due to it comes up with the precise formulation of problem solving and using the right data [17][18][19][20][21]. This paper is about the adaptability of commonly used data mining algorithms through multiagent system in distributed system.

The rest of the paper is organized as: Section 2 is the related literature, section 3 discusses the methodology, section 4 is about results and discussion and finally the conclusion is drawn in section 5.

2. Related Literature

In this section we discuss the most commonly used data mining algorithms and modern distributed computing systems, their issues and applications.

2.1. Data Mining Algorithms

The users may think that only by loading the computer software; data mining will happen. Before moving forward with data mining many issues must be considered such as data pre-processing. There is often a misconception that data mining is a data warehousing, SQL queries and reporting, software agents and online analytical processing (OLAP). The answer is; these are not data mining. Data mining in fact increases computing power, improves data collection and management and it has statistical and learning algorithms. It is clear that decisions are not made by data mining; the people have to decide with their knowledge and experience. The main properties of data mining algorithms are robustness, scalability and efficiency. Data mining is now used in bioinformatics, genetics, medicine, education & electrical power. The major objectives of data mining are to verify the hypothesis prepared by the user and to discover or uncover new patterns. Classification, Clustering, Regression model and Association rule Learning are the main areas of data mining [15][16]. Figure 1 shows the data mining algorithms.

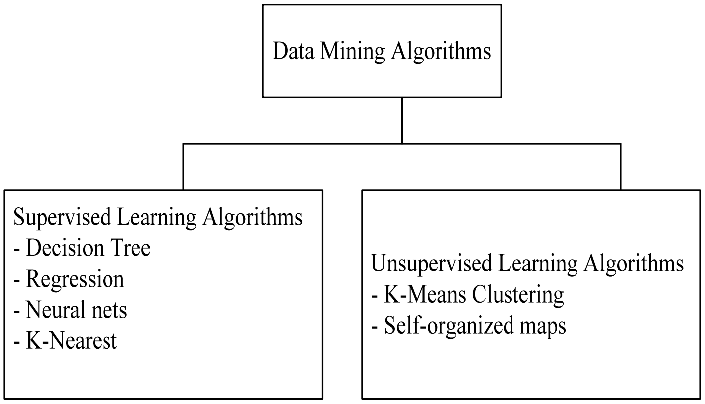

Figure1. Data Mining Algorithms

2.1.1. Neural Networks

Neural networks are used in system performing image and signal processing, pattern recognition, robotics, automatic navigation, prediction and forecasting and simulations. Figure 2 depicts Neural Networks with a hidden layer.

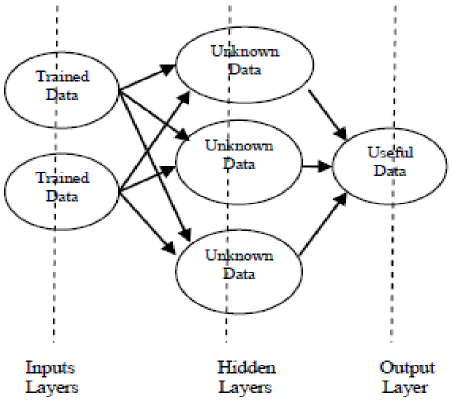

Figure 2. Neural Network with one hidden layer

The issues in NNs are: The process it goes through is considered to be hidden and therefore left unexplained. This lack of explicitness may lead to less confidence in the results and a lack of willingness to apply those results from data mining, since there is no understanging of how the results came about. It is abovious, as the number of variables of a dataset increases, it will become more difficult to understand how the NNs come to it conclusion.The algorithm is better suited to learning on small to medium sized datasets as it becomes too time inefficient on large sized datasets [1][2][3].

2.1.2. C4.5

C4.5 (Decision tree) is used to produce classifiers, which will predict with the values of its available input attributes and the class for an entity. The classification is the process of dividing the samples into pre-defined groups. It is used for decision rules as an output. In order to do mining with the decision trees, the attributes should have continuous discrete values, the target attribute values must be provided in advance and the data must be sufficient so that the prediction of the results will be possible. Decision trees are faster to use, easier to generate understandable rules and simpler to explain. They also help to form an accurate, balanced picture of the risks and rewards that can result from a particular choice. The decision rules are obtained in the form of “if-then-else”, which can be used for the decision support systems, classification and prediction. Figure 3 illustrates how decision rules are obtained from decision tree algorithm.

Data
C4.5 Algorithm
Decision Rules

Figure 3. Function of C4.5 Algorithm

The issues in C.5 are: It is good for small problems but quickly becomes cumbersome and hard to read for intermediate-sized problems. Special software is required to draw that tree. If there is a noise in the learning set, it will fail to find a tree. The data must be interval or categorical. Any data not in this format will have to be recorded to this format. This process could hide relationships. Over fitting, large set of possible hypotheses, pruning of the tree is required. C4.5 generally represents a finite number of classes or possibilities. It is difficult for decision makers to quantify a finite amount of variables. This sometimes affects the accuracy of the output, hence misleading answer. If the list of variables increases the if-then statements created can become more complex. This method is not useful for all types of data mining, such as time series [1][2][3].

2.1.3. K-means Clustering

Unsupervised Learning depends on input data only and makes no demands on knowing the solution. It is used to recognize the similarities between inputs or to identify the features in the input data. It is used for finding the similar patterns due to its simplicity and fast execution. It starts with a random, initial partition and keeps re-assigning the samples to clusters, based on the similarity between samples and clusters, until a convergence criterion is met. The basic working of all the clustering algorithms is represented in figure 4.

Dataset
K-means Algorithm
Clusters of Dataset

Figure 4. Function of the K-means Algorithm

The issues in K-means are: The algorithm is only applicable to datasets where the notion of the mean is defined. Thus, it is difficult to apply to categorical data sets. There is, however, a variation of the *k*-means algorithm called *k*-modes, which clusters categorical data. The algorithm uses the mode instead of the mean as the centroid. The user needs to specify the number of clusters *k* in advance. In practice, several *k* values are tried and the one that gives the most desirable result is selected. The algorithm is sensitive to outliers. Outliers are data points that are very far away from other data points. Outliers could be errors in the data recording or some special data points with very different values. The algorithm is sensitive to initial seeds, which are the initially selected centroids. Different initial seeds may result in different clusters. Thus, if the sum of squared error is used as the stopping criterion, the algorithm only achieves local optimal. The global optimal is computationally infeasible for large data sets. The algorithm is not suitable for discovering clusters that are not hyper-ellipsoids or hyper-spheres [1][2][3].

2.1.4. Genetic Algorithms (GAs)

They are used in optimization problems, selection, crossover and mutation. They are called natural selection and evolution of the problem. The issue in GAs is: It is not used in the large scale problems. It requires a significant computational effort with respect to other methods with parallel processing is not employed [1][2][3].

2.1.5. Fuzzy Logic (FL)

In contras to binary, multi valued logic to deal with imprecise or vague data. It is a new field not very widely used. FL is a fad. There is no guarantee that it will work under all circumstances. The issues in FL are: The language barrier is the major problem. In Japanese term it is ‘clever’ but in American it is ‘fuzzy’. This technology is still somewhat under developed in United States. It seems as if many American researchers have shunned it. It is not yet popular in data mining so it may not get the status due to its name [1][2][3].

2.1.6. Data Visualization

This method provides the user better understanding of data. Graphics and visualization tools better illustrate the relationship among data and their importance in data analysis cannot be overemphasized. The distributions of values can be displayed by using histograms or box plots. 2D or 3D scattered graphs can also be used. [3]Visualization works because it provides the broader information as opposed to text or numbers. The missing and exceptional values from data, the relationships and patterns within the data are easier to identify when graphically displayed. It allows the user to easily focus and see the patterns and trends amongst data. The major issues in data visualization are: As the volume of the data increases it becomes difficult to distinguish patterns from data sets. Another problem in using visualization is displaying multi-dimensional or multi-variable models because only two-dimensions can be shown on a computer or paper [1][2][3].

2.1.7. K-Nearest Neighbor (K-NN)

It is a classification technique that considers the solutions of similar problems that have been solved previously. It decides where to place a new case “k” after examining the most similar cases of “k” or its neighbors. For example “N” is a new case; it is assigned to the class “x”, because the algorithm assigns new case on the bases of most similar cases of “N” or its neighbor. It is illustrated in figure 5 [1][2][3].

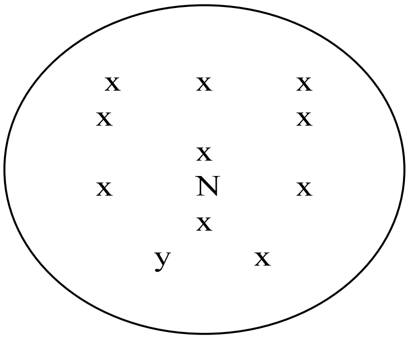

Figure 5. K-Nearest Neighbor

K-NN models can be used in modeling of non-standard data type i.e. text and these models are suitable for few predictor variables because the output is easy to understand. The first thing is to calculate the distance between attributes in the data. Once the distances between cases is calculated, then select the set of already classified cases then decide the range of neighborhood to do comparison also count the neighbors themselves. The major issues in K-NN are: It requires large computational load on the computer therefore all the data is kept in memory. This enhances the speed of K-NN. It is also known as K-NN Memory-based reasoning. It processes new cases rapidly as compared to d. tree or neural networks. It requires new calculation for new cases. Numeric data can easily be handled by this algorithm, categorical variables need special handling [2][3][7].

2.1.8. Bayesian Classification

As its name implies, Bayesian classification attempts to assign a sample x to one of the given classes using a probability model defined according to the Bayesian theorem. The latter calculates the posterior probability of an event, conditional on some other event. Basic prerequisites for the application of Bayesian classification are: Knowledge of the prior probability for each class. Knowledge of the conditional probability density function for each class. It is then possible to calculate the posterior probability using the Bayesian formula

$$q(ci \mid x) = (p(x \mid ci)p(ci)) / p(x) \qquad (1)$$

where p(x) is the prior probability of x. Each new data tuple is classified in the class with the highest posterior probability. The issue in Bayesian Classification is: Major drawbacks of Bayesian classification are the high computational complexity and the need for complete knowledge of prior and conditional probabilities [4].

2.2. Distributed Computing Systems

The goal of distributed computing is to solve large computational problems. These are popular due to two main reasons: First, the nature of the problem requires using communication network that connects several computers and second the use of distributed system is beneficial for practical reasons. The distributed computing provides remote communication, fault tolerance and remote information access. Figure 6 below shows different forms of distributed computing:

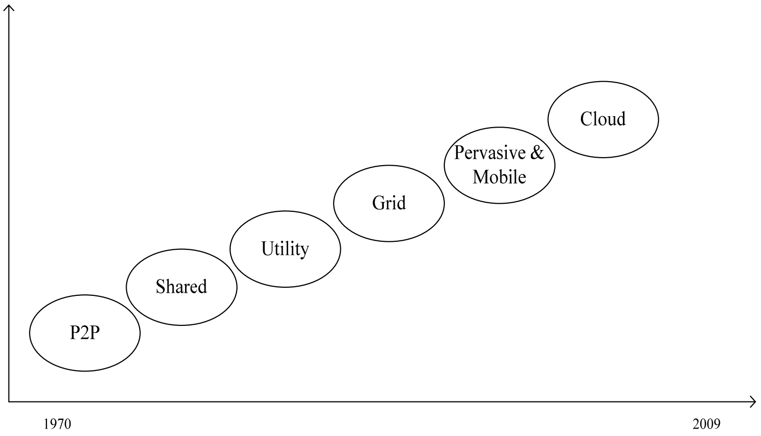

Figure 6. Forms of Distributed Computing

2.2.1. Cloud Computing

It is a new type of distributed computing still emerging field in computer science. The remote machines owned by other company which will run everything for the user is called cloud computing. It will change entire computer industry. The only thing the user has to run interface software of cloud computing system. There is a significant workload shift i.e. the user's computer will not run the applications. It will decrease the demand of hardware and software. There is no limit of its applications. Everything can be done through cloud computing. The major advantage of this is the client can access his data any where at any time. It will reduce the need for advanced hardware which will bring the cost of hardware down. The client can take the advantage of network processing power if the cloud computing is using a grid at its back end. This is a step backward to early computers having only keyboard and terminal. The major issues in cloud computing are: Data governance: Enterprises have sensitive data that requires proper monitoring and protection, moving data into cloud enterprise will lose their governance on own data. Manageability, Monitoring, Compliance, Cross-country Data migration, Reliability, availability and recovery and Security and privacy: are major concerns and issues in cloud computing [7].

2.2.2. Shared Computing

It is a network of computers that work together to complete a task. It is sharing of processing power and other resources. A user can access the processing power of entire network. It is used only for the complex problems not for others. Its administration and design is complicated. The main issues are: The safety and privacy is issue in shared computing. It needs a plan when a system goes offline or unavailable. Power consumption in shared computing is high which produces the heat. The major concern about shared computing is that they are not comprehensive. They uses only processing power not the other resources like storage. The grid computing is more applicable then shared because of its resource sharing [7].

2.2.3. Utility Computing

It is a business model in which a company outsources its computer support to other company. This support can be in the form of processing power, storage, hardware and software applications. The major advantage of utility computing is convenience because client has not to buy all hardware and licensed software for his business. He has to rely on another party to provide these services. The main issues are: This type of computing model is suitable for medium or large scale enterprise, not suitable for small business. Another main disadvantage of utility computing is reliability i.e. clients may hesitate to hand over duties to a smaller company where they feel the lost of data. It is an easy target of hackers. The major challenge in utility computing is that the consumers are not educated about its service. Its awareness is not very widespread [5].

2.2.4. Grid Computing

It is a type of distributed computing where every computer can access the resources such as processing power, memory and data storage of other computer on the network, turning it into a powerful supercomputer. It is a high performance computing. The grid computing can provide an effective computational support from applications for knowledge discovery. The basic services of grid computing are communication, authentication, information and resource management. This is not a new concept but not yet perfect. People are still working on creating, establishing and implementing standards and protocols. The applications of grid computing are limitless. The main challenges are: Resource sharing & coordinated problem, Coordinated problem solving in dynamic, Multi-institutional virtual organizations, Data protection, No clear standard, Better understanding as simple as possible, Difficult to develop, Lack of grid-enabled software, Centralized management, and the limited number of users are allowed the full access of network otherwise the control node will be flooded with processing requests which can create deadlock situation [10][13].

2.2.5. Pervasive Computing

It is a share of small, inexpensive, robust networked processing devices distributed at all scales in everyday life. The main challenges in pervasive computing are: System design and engineering, System modeling and human computer interaction models [9].

2.2.6. Mobile Computing

It is an ability to use technology while moving. It is a building block for pervasive computing. It is a tomorrow's network technology. It will revolutionize the way computers are used. Wireless communication, mobility and portability are salient features of mobile computing. It is a paradigm shift in distributing computing. The limitations of mobile computing are: There are no such standards of security of data. The time of power supply is limited. There may be signal problem due to transmission interferences. The excess use of mobile devices may cause potential health hazards. Small and limited user interface for human on the device is available. Small storage capacity, Risks to data, Mobility, moving from one coverage area to another and insufficient bandwidth [7].

3. Methodology

The traditional centralized data analyzing is not suitable for distributed applications. In distributed environment analyzing the distributed data is a non-trivial problem because of many constraints such as limited bandwidth, privacy-sensitive data and distributed nodes. The distributed problems solving environment fit well with the multiagent system (MAS) since the solution requires autonomous behavior, collabartion and reasioning. The agents perform the underlying data analysis task very efficiently in distributed manner. The MAS offer an architecture for distributed problem solving, deal with complex applications that require distributed problem solving. The combination of data mining algorithms and MAS for data analyzing will further enhance the processing power of the application.

Two medical datasets 'Diabetes' and 'Breast Cancer' are selected for this paper. The data is pre-processed, called data standardization. The interval scaled data is properly cleansed by applying the *range method*. The attributes of the dataset/testbed 'Diabetes' are: Number of Times Pregnant (NTP)(min. age = 21, max. age = 81), Plasma Glucose Concentration a 2 hours in an oral glucose tolerance test (PGC), Diastolic Blood Pressure (mm Hg) (DBP), Triceps Skin Fold Thickness (mm) (TSFT), 2-Hour Serum Insulin (m U/ml) (2HSHI), Body Mass Index (weight in kg/(height in m)^2) (BMI), Diabetes Pedigree Function (DPF), Age, Class (whether diabetes is cat 1 or cat 2) [22][23][24][25]. The attributes of dataset 'Breast Cancer' are: Clump Thickness (CT), Uniformity of Cell Size (UCS), Uniformity of Cell Shape (UCSh), Marginal Adhesion (Mad), Single Epithelial Cell Size (SECS), Bare Nuclei (BNu), Bland Chromatin (BCh), Normal Nucleoli (NNu), Mitoses , Class (benign, malignant). The vertical partitions of the chosen datasets are created. The architecture of MAS is shown in figure 7.

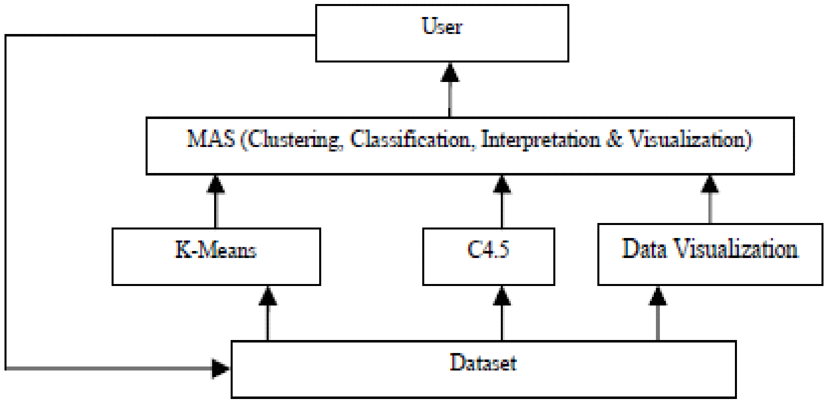
User
MAS (Clustering, Classification, Interpretation & Visualization)
K-Means
C4.5
Data Visualization
Dataset

Figure 7. The Architecture of MAS

The partitioned datasets are used as inputs, three intelligent agents for conventional data mining algorithms such as k-means, c4.5 and data visualization, take these datasets and produce the outputs clusters, decision rules and 2D graphs which are helpful for users to take decisions. The MAS is deployed on a client-server based local distributed system in the form of a 'grid computing' shown in figure 8.

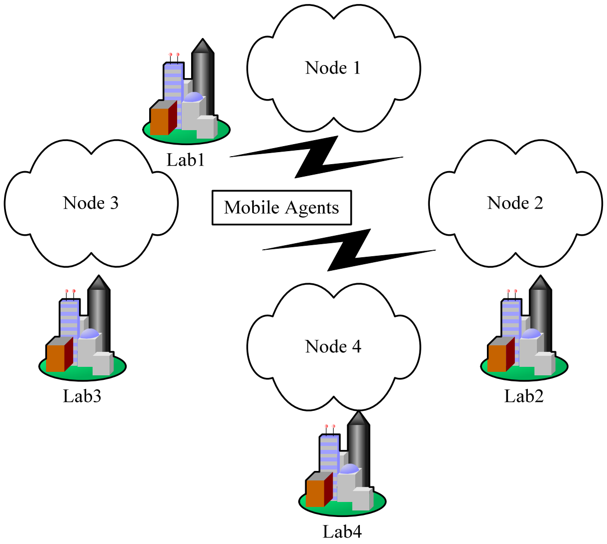

Figure 8. A distributed network using MAS

These agents can roam from one node to other node freely and can be stored at any node in the distributed network. The results of these agents can be stored at any where in the network. The user has the access to all the outputs [11][12][14].

4. Results and Discussion

The results and discussion part of this paper has two parts. In the first part, we draw a comparison of commonly used data mining algorithms in table 1. In order to build a predictive model, understanding of the data is the must. The data is in the forms of continuous and categorical (ordinal or nominal).

Table 1. Comparison of Data Mining Algorithms

Properties	**NNs**	**DTs**	**K-means**	**FL**	**GAs**	**Data visualization**	**K-NN**	**Bayesian Classification**
Training/Learning of a data set	High	Low	Low	None	High when no parallel processing	None	High	Low
Explicitness	Yes	No	No	None	No	No	No	No
Lack of Knowledge	Yes	No	Yes	It is a fad	No	No	No	No
Lackness of problem solving	Yes	Yes	Yes	None	No	Yes	Yes	No
Type of data in data sets	All types of data sets	Interval or categorical	Categorical and Numeric	None	It is suitable for strings	All types of data used in graphs	Numeric	Numeric
Estimation or prediction	Yes	No estimation but prediction is possible	Yes	Not popular in data mining	General-purpose search algorithm	Yes	Yes	Yes
Resource consumption	High	Medium	Medium	None	Low	Low	High	High
Size of data sets	Small to Intermediate	Small	Small to intermediate	None	Small	Small to Intermediate	Small to Intermediate	Small to Intermediate
Complexity	O(m)	O(m)	O(nkl) for time and O(k+n) for space	O(m)	O(m)	O(m)	O(m)	O(m)

The main properties of data mining algorithms are robustness, scalability and efficiency. It is clear from this comparison that all the data mining algorithms are suitable and produce useful results only on small to medium scale datasets. They are not appropriate for large datasets. This is a problem of scalability. There is another overhead; the learning and training of the dataset is required in all algorithms. If there is any noise in datasets, the extracted knowledge may be misleading. The choice of the algorithm depends on the intended use of extracted knowledge. The data can be used either to predict future behavior or describe patterns in an understandable form within discover process.

We also discuss different distributed systems and their issues. The induction of cloud computing has change the paradigm; it is a step backward to early computer having only keyboard and terminal. The distributed systems provide remote access to information, when resources are shared in any type of distributed computing; the users have to compromise on privacy and security of their data. These are two major issues in all form of distributed computing. There are no standards in design, implementation and management of networks in any form of distributed computing. The migration of data across the country is another issue. A mountain of data is there and to fetch the correct data on the request of a user from the clouds of data is another issue.

In the second part, the results obtained from the agents of the MAS are discussed. The output of these agents is in the form of 2D graphs of the clusters, 2D graphs of whole partitioned clustered dataset and decision rules in the form of if-then-else. We discuss some of the graphs and decision rules in this paper. The cluster by cluster graphs and the whole partitioned clustered graphs have their own significance and the user can use either of the graphs or both for the future prediction of the data. The purpose of 2D graph is to identify the type of relationship if any, between the attributes of the given dataset. The graph is used when a variable exists which is being tested and in this case the attribute or variable ‘class’ is a test attribute of both selected medical datasets.

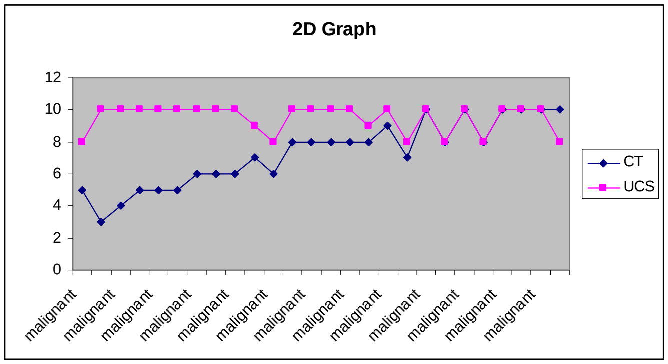

Figure 9. 2D graph between attributes ‘clump thickness’ & ‘uniformity of cell size’ of dataset ‘Breastcancer’

The graph in figure 9 shows that the value of the attributes ‘clump thickness’ and ‘uniformity of cell size’ is almost constant at the beginning and at the end of the graph both attributes have equal value. The outcome of this graph is that if the value of these attributes is either constant or equal then the patient has ‘malignant’ class of breast cancer.

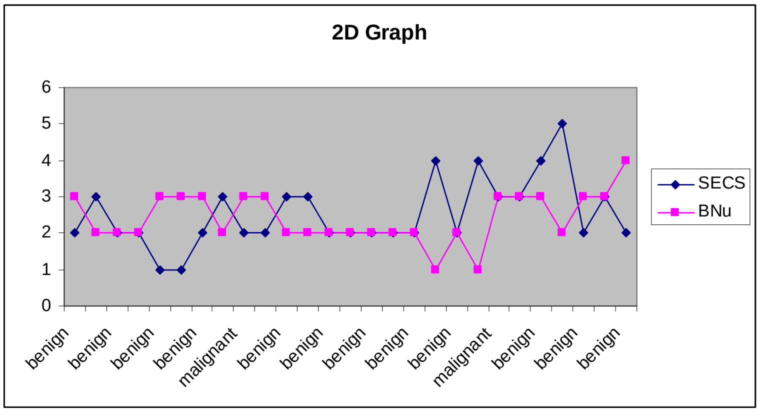

Figure 10. 2D graph between attributes ‘single epithelial size’
& ‘bare nuclei’ of dataset ‘Breastcancer’

The structure of graph in figure 10 is almost similar in figure 9. This graph has three sections, the value of attributes ‘single epithelial size’ and ‘bare nuclei’ is constant in the beginning, then the value of these attributes is variable, then in some region of the graph the value is equal and then again it is constant. The conclusion of this graph is that if the value of these attributes is either constant or equal then the patient has ‘benign’ class of breast cancer and the variable values of these attributes give the ‘malignant’ class of breast cancer.

Figure 11 shows the decision rules of a cluster of dataset ‘Breast Cancer’.

1. Rule: 1 if Mitoses = 1 then
2. Rule: 2 if BCh = 7 then
3. Class = malignant , benign else
4. Rule: 3 if BCh = 1 then
5. Class = benign else
6. Rule: 4 if BCh = 2 then
7. Class = benign , malignant else
8. Rule: 5 if BCh = 3 then
9. Class = benign , malignant else
10. Rule: 6 if BCh = 4 then
11. Class = malignant , benign else
12. Rule: 7 if BCh = 8 then
13. Class = malignant

Figure 11. The Decision Rules of a cluster of dataset ‘Breast Cancer’

The decision rules in the form of ‘if – then – else’ are shown. It is an easy way to take the decision. The result is if the value of the attribute ‘mitoses’ is 1 and the value of attribute ‘bland chromatin’ is either 1, 2 or 3 then the patient has ‘benign’ class of breast cancer otherwise the ‘malignant’ class of breast cancer. These rules can be used in a simple query to further validate the results.

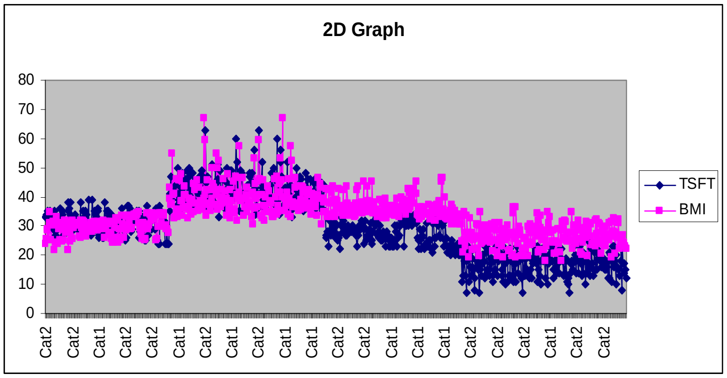

Figure 12. 2D graph between attributes 'triceps skin fold thickness' & 'body mass index' of dataset 'Diabetes'

The graph of the whole partitioned clustered dataset 'Diabetes' is shown in figure 12. The structure of the graph is complex as compared to the graphs of 'Breast Cancer' dataset. The graph has two regions, either the value of attributes 'triceps skin fold thickness' and 'body mass index' is equal or constant. There is no variation in values of the attributes in this graph. It is difficult to take the decision either the patient has 'cat1' or 'cat2 type of diabetes.

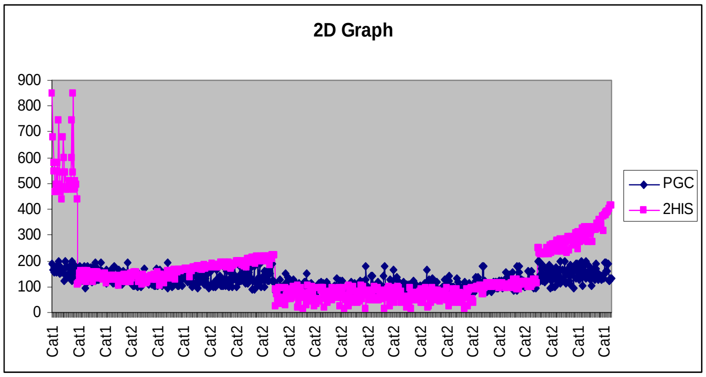

Figure 13. 2D graph between attributes 'plasma glucose concentration' & '2-hour insulin serum' of dataset 'Diabetes'

The graph of the whole partitioned clustered dataset 'Diabetes' is shown in figure 13. The graph has two main regions, in the beginning and at the end the value of attributes 'plasma glucose concentration' and '2-hour insulin serum' is variable, which shows the 'cat1' type of diabetes in patient and in the middle of the graph the value of these attributes is either constant or equal, which gives 'cat2' type of diabetes in patient.

Figure 14 shows the decision rules of whole partitioned clustered dataset 'Diabetes'.

1. Rule: 1 if PGC = 165
   then
2. Class = Cat2
   else
3. Rule: 2 if PGC = 153
   then
4. Class = Cat2
   else
5. Rule: 3 if PGC = 157
   then
6. Class = Cat2
   else
7. Rule: 4 if PGC = 139
   then
8. Class = Cat2
   else
9. Rule: 5 if HIS = 545
   then
10. Class = Cat2
    else
11. Rule: 6 if HIS = 744
    then
12. Class = Cat2
    else
13. Class = Cat1

Figure 14. The Decision Rules of whole partitioned clustered dataset 'Diabetes'

The result for this whole partitioned clustered dataset 'Diabetes' is if the value of attribute 'plasma glucose concentration' is more than 130 and the value of attribute '2-hour insulin serum' is more than 500, as the rules show, then the patient has diabetes of cat2 otherwise cat1. The decision rules make it easy and simpler for the user to interpret and predict this partitioned dataset of diabetes. These rules can be used in a simple query to further validate the results.

Similarly, the 2D graphs and the decision rules can also be drawn for the other clusters and partitioned clustered datasets. These results show that the conventional data mining algorithms can be deployed using the MAS on the modern distributed systems.

5. Conclusion

The goals of data mining are pattern and feature extraction, visualization of data and evaluation of results. These goals can be achieved by using different data mining algorithms. In this paper we discuss the most commonly used data mining algorithms, their issues and applications and a comparison of these algorithms is drawn. If the user is clear about business goals, type of prediction and model then the selection of algorithms is very easy. Three data mining 'k-means clustering', 'C4.5' and 'data visualization' algorithms are selected, for clustering, classification, interpretation and visualization. A MAS approach comprising three intelligent agents, one for each algorithm is used. Two medical datasets 'Breastcancer' and 'Diabetes' are selected to test the MAS over a client-server based local distributed system in the form of a 'grid computing'. The vertical partitions of the given datasets, based on the similar values of the attributes are created. The results obtained are encouraging, acceptable, satisfactory and consistent. We conclude our paper, due to wide spread use of distributed systems nowadays, in order to access the data and produce 'knowledge', apply intelligent mobile agents in data mining algorithms over distributed systems.

Future Work

The study could be extended to large scale distributed systems with the other data mining algorithms so as to validate the effectiveness of the proposed methodology. For further investigation in this direction, one will undoubtedly has to take into account the parameters such as data caching and the validity of the agent framework.

Acknowledgment

The authors are thankful to ‘The Islamia University of Bahawalpur, Pakistan’ for providing financial assistance to carry out this research activity under HEC, Pakistan project 6467/F – II.

References

[1] Wang, John., *“Data Mining Opportunities and Challenges”*, Idea Group Publishing ISBN: 1-59140-051-1, chapter IX page 235 and chapter XVI page 381

[2] Liu, Bing., “*Web Data Mining Exploring Hyperlinks, Contents, and Usage Data”*, ISBN: 13 978-3-540-37881-5, Springer Berlin Heidelberg New York, chapter 3 and chapter 4.

[3] *“Introduction to Data Mining and Knowledge Discovery”*, ISBN: 1-892095-02-5, Third Edition by Two Crows Corporation, page numbers: 11,12,13,15.

[4] Symeonids, Andreas. Pericles, Mitkas., *“AGENT INTELLIGENCE THROUGH DATA MINING”*, ISBN 0-387-24352-6, chapter 1, 2, 3.

[5] Y. Chee, B. Rajkumar, D. Marcos, Y. Jia, S. Anthony, V. Srikumar, P. Martin, “*Utility Computing on Global Grids”*, ISBN: (978-0-471-78461-6, John Wil ey & Sons, New York, USA, 2007.

[6] L. Jiangchuan, Rao. Sanjay, Li. Bo, Z. Hui, “Opportunities and Challenges of Peer-to-Peer Internet Video Broadcast”, *Proceedings of the IEEE* Volume 96, Issue 1, Jan. 2008 Page(s):11 - 24

[7] Satyanarayanan. M., “Fundamental Challenges in Mobile Computing”, School of Computer Science, Carnegie Mellon University, *Proceedings of the fifteenth annual ACM symposium on Principles of distributed computing* Philadelphia, Pennsylvania, United States, Pages: 1 – 7, ISBN:0-89791-800-2

[8] Jimenez. Raul, Eriksson. Lars-Erik, Knutsson, Bj¨orn., “P2P-Next: Technical and Legal Challenges”, *6th Swedish National Computer Networking Workshop (SNCNW'09) and 9th Scandinavian Workshop on Wireless Adhoc Networks (Adhoc'09) Sponsored by IEEE VT/COM Sweden*

[9] Satyanarayanan. M., “Pervasive Computing: Vision and Challenges”, *School of Computer Science, Carnegie Mellon University, Personal Communications, IEEE* Aug 2001, Volume: 8, Issue: 4 On page(s): 10-17, ISSN: 1070-9916

[10] C. Mario., T. Domenico., T. Paolo., “Distributed data mining on the grid”, *ICAR-CNR*, Via P. Bucci, Cubo 41-C, 87036, Rende (CS), Italy, August 2002.

[11] Voskob. Max., Howey. Rob., Panin. Nick., “Data mining and Privacy in Public Sector using Intelligent Agents”, *Discussion paper*, November 2003, Wellington, New Zealand.

[12] Grossman. Robert, Kasif. Simon, Moore. Reagan, Rocke. David, Ullman. Jeff, “Data Mining Research: Opportunities and Challenges”, *A Report of three NSF Workshops on Mining Large, Massive, and Distributed Data*, 1998.

[13] Dutta. Haimonti., “Empowering Scientific Discovery by Distributed Data Mining on the Grid Infrastructure”

[14] da Silva. C. Josenildo., et al, “Distributed Data Mining and Agents*”*

[15] Berkhin, Pavel., “*Survey of Clustering Data Mining Techniques”*, Accrue Software, Inc.

[16] Freitas, Alex A., “*A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery”*, Pontificia Universidade Catolica do Parana Rua Imaculada Conceicao, 1155. Curitiba - PR. 80215-901. Brazil.

[17] Mento, Barbara. and Rapple, Brendan., “*Data Mining and Data Warehousing”*, Virtual Data Center Manager Boston College, July 2003

[18] Mangina, Eleni., “Intelligent Agent-Based Monitoring Platform for Applications in Engineering”, *International Journal of Computer Science & Applications* Vol. 2, No. 1, pp. 38 – 48, Technomathematics Research Foundation, 2005.

[19] Chen, Ming-Syan, Han, Jiawei and Yu, Philip S., “Data Mining: An Overview from Database Perspective”, *IBM T.J. Watson* Res. Ctr. P.O.Box 704, Yorktown, NY 10598, U.S.A.

[20] Maindonald, John., “Data Mining Methodological Weaknesses and Suggested Fixes”, *Proc. Fifth Australasian Data Mining Conference (AusDM2006)*, 2006.

[21] Schommer, Christoph., “An Unified Definition of Data Mining”, *arXiv:0809.2696v1 [cs.SC]* 16 Sep 2008.

[22]US Census Bureau. Iris, Diabetes, Vote and Breast datasets at URL: www.sgi.com/tech/mlc/db visited 2009.

[23]National Institute of Diabetes and Digestive and Kidney Diseases, Pima Indians Diabetes Dataset, http://archive.ics.uci.edu/ml/datasets/Diabetes, visited 2009

[24]Skrypnik, Irina., Terziyan, Vagan., Puuronen, Seppo., and Tsymbal, Alexey, “Learning Feature Selection for Medical Databases”, *CBMS 1999*.

[25] Irene M. Mullins et al., “Data mining and clinical data repositories: Insights from a 667,000 patient dataset”, *Computers in Biology and Medicines* 36 p. 1351 – 1377, 2006.