CADAR: A KNOWLEDGE ANALYZER USING CELLULAR LEARNING AUTOMATA AND DISTRIBUTED CELLULAR LEARNING AUTOMATA

MANSOUR ESMAEILPOUR

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2012

DECLARATION

I hereby declare that this research is my own work and effort. It has not been submitted to any other university.

6. September. 2012

MANSOUR ESMAEILPOUR

P47843

ACKNOWLEDGMENT

I would like to thank my supervisor, Prof. Dr. Zarina Shukur and my co-supervisor Dr. Shahnorbanun Sahran, for the patient guidance, encouragement and advice they have provided throughout my time as them student. I have been extremely lucky to have a supervisor who cared so much about my work. I would also like to thank Universiti Kebangsaan Malaysia (UKM) and specially Faculty of Information Science and Technology that sponsored me in during my study.

I need to thank Vahideh Naderifar, my wife, for encouraging and helping me to do this research and I would also like to acknowledge all the members of staff at Hamedan Branch, Islamic Azad University, Hamedan, Iran and department of computer engineering.

CADAR: PENGANALISA PENGETAHUAN OLEH AUTOMATA PEMBELAJARAN BERSEL DAN AUTOMATA PEMBELAJARAN BERSEL TERAGIH

ABSTRAK

Analisis pengetahuan merupakan satu pendekatan untuk menganalisis masalah, isu atau peristiwa melalui perspektif pengetahuan untuk memahami mereka pada tahap yang lebih mendalam dan membentuk pandangan baru tentang mereka. Dalam karya ini, keluarga automata pembelajaran iaitu automata pembelajaran bersel (CLA) dan automata pembelajaran teragih (DLA) digabungkan untuk mendapatkan masa pemprosesan dan ketepatan yang lebih baik dalam menganalisa pengetahuan. Teknologi dibangunkan untuk menyokong keseluruhan proses analisa menggunakan kaedah hibrid ini. Nama teknologi ini adalah Cadar. Secara individu, CLA dan DLA adalah kaedah yang terbukti berkeupayaan tinggi untuk menganalisa pengetahuan seperti pengelompokan, pengelasan, ramalan dan perlombongan corak yang kerap. Namun berdasarkan literatur, tiada kajian yang telah dilakukan untuk membuktikan keupayaan menggabungkan CLA dan DLA dalam menganalisa pengetahuan. Cadar pula dibangunkan dengan tujuan untuk menyediakan persekitaran yang kondusif bagi penganalisa pengetahuan untuk menggunakan kaedah yang dicadangkan. Dari segi konsep, automata pembelajaran (LA) terdiri daripada dua bahagian; automata stokastik dengan beberapa tindakan yang terhad dan persekitaran stokastik. Setiap tindakan yang dipilih oleh persekitaran yang berpotensi akan dinilai dan jawapan akan diberikan kepada automata pembelajaran. CLA pula terdiri daripada dua model; automata pembelajaran dan automata selular. Automata pembelajaran diumpukkan kepada setiap sel dalam automata selular. Ia sesuai untuk digunakan bagi sistem yang boleh diwakili oleh model selular yang mana tingkah laku setiap sel adalah berdasarkan tingkah laku jirannya dan pengalaman lalu. Manakala DLA pula adalah rangkaian pembelajaran automata yang bekerjasama untuk menyelesaikan suatu masalah. Bagi kaedah yang dicadangkan ini, setiap nod daripada DLA adalah satu CLA. Nod dicipta secara dinamik berdasarkan data. Intipati kaedah hibrid ini adalah keupayaan untuk mencerap jumlah tetulang hubungan yang kukuh. Algoritma bagi kaedah yang dicadangkan ini adalah enjin utama Cadar. Ia dilaksanakan dengan menggunakan tatasusunan dinamik yang mana saiznya boleh berubah sewaktu larian. Dengan adanya tatasusunan dinamik, ia akan meningkatkan kecekapan masa pemprosesan fizikal. Selain algoritma ini, Cadar juga mempertimbangkan aspek kebolehgunaan supaya penganalisa pengetahuan boleh menggunakan teknologi dengan lebih baik. Prestasi model yang dicadangkan berbanding kaedah terkenal lain ditentukan dengan menggunakan experimen. Hasilnya menunjukkan bahawa model yang dicadangkan boleh memproses semua jenis dataset dan ia sesuai dan boleh diterima dari aspek masa latihan dan cemerlang dalam ketepatan berbanding dengan kaedah lain. Bagi teknologi pula, hasilnya menunjukkan bahawa Cadar mengatasi WEKA atas empat fungsi: pembersihan data dan pendiskritan, laporan pengelompokan data bergrafik, fungsi perlombongan pola yang kerap dan penjana peraturan.

CADAR: A KNOWLEDGE ANALYZER USING CELLULAR LEARNING AUTOMATA AND DISTRIBUTED CELLULAR LEARNING AUTOMATA

ABSTRACT

Knowledge Analysis is an approach to analyze problems, issues or events through a knowledge perspective in order to understand them at a deeper level and form new insights about them. In this work, a family of learning automata that is Cellular Learning automata (CLA) and Distributed Learning automata (DLA) are blended together in order to obtain a better processing time and accuracy in analyzing knowledge. A technology is developed in order to support the overall process of the analysis using this hybrid method. The name of this technology is *CADAR*. Individually, CLA and DLA are proven powerful methods for knowledge analyzing such as clustering, classification and frequent pattern mining. However, based on the literature, no work has been done to demonstrate the ability of combining CLA and DLA in analyzing knowledge. It is hypothesized that by combining both methods, better result especially in accuracy and run time can be obtained. As for *CADAR*, its aim is to provide a favorable environment for knowledge analyst to use the proposed method. Conceptually, learning automata (LA) is composed of two parts; stochastic automata with a number of limited actions and a stochastic environment. Each action selected by potential environment is assessed and answer is given to a learning automata. LA uses this answer in order to select its action for the next stage. As for CLA, it is composed of two models; learning automata and cellular automata. The learning automata is assigned to every cell in cellular automata. It is suitable to be used for systems that can be represented by a cellular model where each cell's behavior is based on its neighbor's behavior and past experience. Whilst DLA is a network of learning automata which cooperates with each other for solving problem. In this research, in order to acquire better result, it is combined CLA and DLA for knowledge analysis. In this proposed method, each node of the DLA is one CLA. The node is created dynamically based on the data. The gist of this hybrid method is the ability to capture the amount of strong relationship reinforcement. Algorithm of the proposed method is the main engine of *CADAR*. It is implemented by using dynamic arrays where its size can change during runtime. By having dynamic arrays, it increases the efficiency of processing time. Besides the algorithm of the methods, *CADAR* also consider usability aspects so that the knowledge analyst can have better use of the technology. The performance of the proposed method compared to well known methods have been investigated by using an experiment. Whilst the performance of the technology itself compared to WEKA tool is done by using a simple usability study. The result demonstrates that the proposed method can works on all of kind of datasets and it is suitable and acceptable from the aspect of the run time and excellent in accuracy compared to other methods. As for the technology, the result shows that *CADAR* wins over WEKA on four features that are: data cleaning and discretization function, graphical report of data clustering, frequent pattern mining function, and rule generator.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

This chapter contains a short introduction of the topic of the thesis as well as a description of what problem the thesis wants to answer. The goal, purpose and target group, who will benefit from the project, will be presented here. Any limitations of the project will be coverd in the last part of this chapter.

First of all, the analysis of the knowledge and extraction of the valid relationship in the datasets are important with respect to expanding the size of useful data. So the use of a suitable method in analyzing the knowledge leads to a decrease in errors and an increase in accuracy and performance. Meanwhile, the intelligent gain knowledge is understanding, or skill by the study, experience and modification of a behavioral tendency. In other word, knowledge analysis is aimed at studying knowledge intensive tasks at a conceptual level. The analysis results in a description of the information and knowledge structures and functions involved in the task.

There are several ways to analyze the knowledge such as data mining, time-series analysis; benchmarking, etc. Data mining is a process to analysis step of the knowledge discovery in datasets. The goal of data mining is to extract knowledge from a data set in a human-understandable structure (Chakrabarti et al. 2006). Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Knowledge Analysis is an approach to analyze problems, issues or events through a knowledge perspective in order to understand them at a deeper level and form new insights about them. In order to analyze knowledge, a suitable

algorithm must be selected. Some algorithms and methods are suitable for some data, and otherwise. Therefore, an algorithm should be selected for the data on the basis of special method and each algorithm may announce different accuracy with various performances. Set of feature in dataset represent samples of the data. These features may be continuous, clustered or binary. If samples have been specified with a label, then, it calls supervised learning or otherwise it calls unsupervised learning. Another type of machine learning is the reinforcement learning that this method contains an agent that studies environment which may receive either the reward or penalty.

Generally, many machine learning techniques are derived from the efforts of psychologists to make more precise their theories of animal and human learning through computational model. Most of knowledge analysis algorithms are NP-hard, if one algorithm is efficient for one NP-hard problem, then the algorithm would be applicable to all NP-hard problems. Therefore, this work introduced a method for knowledge analyzing based on reinforcement learning that the accuracy and performances of a proposed method will be compared with other methods.

1.1 BACKGROUND STUDY

Since computer technology has increased the power of ever-expanding data collection, storage and manipulations, several common types of tasks have been presented for knowledge analyzing. These are briefly referred to as follows:

- Data classification. One of the unsupervised learning that is used for predicting and categorizing of data. This can be thought of as two separate problems: binary classification and multi-class classification. In binary classification, only two classes are involved, whereas multiclass classification involves assigning an object to one of several classes (Har-peled et al. 2003, Kotsintis et al. 2006 and Wieland et al. 2008).
- Data clustering. A supervised learning, which is a common technique for statistical data analysis. This type of task is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so

that the data in each subset shares some common trait, often proximity according to some defined distance measure (Didy 1973, Han et al. 2001 and Kao et al. 2008).

- Frequent pattern mining. This algorithm is used for studying about human behavior, events, online activities and etc. Frequency patterns are sets of items, sequences, or infrastructures that could be repeated in a data set or what is determined by the user as a threshold limit. The study of frequency patterns was first developed by Agrawal in 1993 followed by extended ideas of his work (Agrawal et al. 1993 and Grahne and Zhu 2003).

1.2 PROBLEM STATEMENT

Increasing the size of the knowledge in each field and also needs to analyze and extracting the valid relationship in this knowledge, leads to increase the research by the researchers in the fields of data mining.

Data classification, data clustering and frequent pattern mining have been selected in this research. These three tasks are much more important than other data mining tasks, and in recent years have been used widely in the areas of science and engineering. Classification is a supervised learning method as well as, clustering and frequent pattern mining is an unsupervised learning.

Accuracy is an important factor in assessing the success of data mining. When applied to data, accuracy refers to the rate of correct values in the data. When applied to models, accuracy refers to the degree of fit between the model and the data. Running time and training time are very important for real-time instruction detection, internet activity, and banking; and shows the ability of the model from aspect of the time complexity. Error rate is calculated by the predicting results on the training examples, in other word, error rate is percent of misclassified examples. Several researchers try to present new techniques to improve the ability of these tasks from the above aspects but most of these methods suffer the same problem that as bellow:

- Data classification: Decision trees splits based on a single feature at each internal node. Most decision tree algorithms cannot perform well with problems that require diagonal partitioning (Kotsiantis 2007). Another method is naïve bayes classifier that the assumption of independence among child nodes is clearly almost always wrong, due to that naïve bayes classifier has less accurate that other learning algorithms (Kononenko 1991).

- Data clustering: Ant colony and cellular methods usually need high iteration in order to find the clusters as well as there is no mechanism at first in the ants-based methods, which can be regarded as basis for targeted movement of ants in a cell net surface and cause to increase running time and delayed convergence to an answer in clustering (Xu et al. 2005). Other methods need to more time for finding the clusters and also have more error to arrange them.

- Frequent pattern mining: Frequency mining methods try to reach frequent patterns in the acceptable running time. Most of these models suffer from increasing the running time that can referrer to Park and Chen (1995). For example Apriori method needs to *n* time scan on dataset, so that increase the running time of the model. Grahne & Zhu (2003) have provided simplification method on the basis of array but counting the elements in the array takes more time in high density datasets. Using array in dense dataset may not be good idea and also these methods are not suitable for real-time instruction detection for example in the medical and real-time datasets do to need to more scan of dataset.

- Develop an Environment: Algorithms are useful to the end user when it is implemented as a package in a complete set of provision. For example, spreadsheet such as Excel has several functions which work on the data but does not have any function for analyzing the knowledge. As for WEKA (Waikato Environment for Knowledge Analysis) it is collection of machine learning algorithms for solving data mining problems and has several functions to analysis the knowledge (Witten et al. 2005), but WEKA does not support the CLA and DCLA models for knowledge analyzing. Therefore, this research

will also provide an environment which is called CADAR to analyze knowledge using a knowledge analysis model based on CLA and DCLA.

1.3 RESEARCH OBJECTIVE

The general aim of this thesis is to develop a model for knowledge analyzer using cellular learning automata and distributed cellular learning automata.

Specifically, the objectives of this thesis can be stated as:

1. To propose a knowledge analysis model based on the CLA for data classification.
2. To propose a knowledge analysis model based on the DCLA for data clustering.
3. To propose a knowledge analysis model based on the CLA for frequent pattern mining.
4. To develop an environment for analyze the knowledge using a knowledge analysis model based on CLA and DCLA.

1.4 RESEARCH METHODOLOGY

The library research was done in this thesis to understand the principle of cellular learning automata and distributed learning automata as reinforcement learning. Following that the characteristics, weaknesses and strength points of CLA and DLA are extracted. All the tasks of the data mining were studied and three of them were selected due to close to ability of CLA and DLA problem solving. These three tasks are data classification, data clustering and frequent pattern mining.

After understanding the insight principle of them, a new model of the CLA for data classification and frequent pattern mining; and a hybrid model of the CLA and DLA which is called DCLA for data clustering are introduced. In this research, several standard respective online dataset were obtained from the UCI machine learning dataset (UCI repository of machine learning datasets) for each three parts separately.

The proposed models were evaluated based on experimental results from the aspect of accuracy of the models, error rate, running time and training time. This step was done by comparing the proposed model with other related models using WEKA software

(Witten et al., 2005). Finally, an environment was provided as a knowledge analysis software based on CLA and DCLA which is called CADAR.

1.5 POTENTIAL CONTRIBUTION

CLA and DLA are well known approaches for solving problems in application domain and computing domain. Technically, when solving problems in application domain, no changes are done to the CLA and DLA model which one can refer to (Mojaradi et al. 2004**)** and (Hamou et al. 2010). However when solving problems in computing domain, the behaviors or parameters of the CLA and DLA model need to be modified as bellow:

- Re-present new neighborhoods.
- Re-introduce local rules.
- To modify the structure of the model.
- To optimize the behavior of the model.

In this research, several modification and optimization have been done on the model to increase the accuracy and decrease the error rate and running time; and also will introduce a new model of CLA named CLA-classifier for data classifying, a hybrid model of CLA and DLA named DCLA-clustering for achieving the desirable clusters and CLA-FPM for frequent pattern mining.

The proposed models can analyze the knowledge better than other models and the experimental results demonstrate the proposed models are much better than others in terms of accuracy, as well as, will decrease the error rate, running time and training time.

1.6 LIMITATIONS

The findings of this study were limited by the following:

1. The findings of this study can identify connections between each pair wise attributes; it does not necessarily identify a causal relationship.

2. The scope of this study can help reveal patterns, clusters and relationships, it does not tell the user the value or significance of these patterns, clusters and relationships.

1.7 ORGANIZATION OF THESIS

The remainder of this thesis organized as follow:

Chapter II introduces the literature review of the work and Chapter III describes the research methodology, Chapter IV consists of background of CLA and DLA can shows how can use basic CLA and DLA for solving problems. Proposed method will be introduced in Chapter V that show the contribution of the model based on advance CLA and DCLA for data classification, data clustering and frequent pattern mining. Chapter VI introduces experimental results of the proposed method and will be studied implementation of the CADAR as knowledge analyzer software, usability of the CADAR software and comparison between CADAR and WEKA in Chapter VII. Finally, Chapter VIII describes the conclusion of this research.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

There are different learning methods in machine learning of which each has both advantages and disadvantages. Each type provides a better and interesting position, data and special structure. These methods have differences in the manner of implementation, understandability and speed of response; and each is included in a special field of the data mining. The learning process and knowledge analysis are the most important parts in data mining, which cause the elevation of power in a model, so one can learn the trained problem more quickly.

As mentioned in Chapter I, most of knowledge analysis algorithms are NP-hard that several researchers have been trying to find more efficient heuristic algorithms to increase the accuracy; and decrease the errors and running time. As for using heuristic algorithms for knowledge analysis, learning automata, distributed learning automata and cellular learning automata approaches will be represented in this research to achieve the goals. In this research learning automata, distributed learning automata and cellular learning automata approaches are presented to achieve the goals. The following briefly described the three categories of knowledge analyzing techniques:

- **Supervised learning:** This method reaches final results by using sets of acceptable input data and output data. In this method, the data need to be labeled.

- **Unsupervised learning:** This method receives input data and a model then can map the observations with the use of present model.

- **Reinforcement learning**: This method contains an agent and an environment which may receive either reward or penalty. The selected action is applied to the environment. The environment investigates it and reward or penalty is chargeable to environment which will update the model with these values.

In this section discusses about knowledge analysis from three aspect, first section is about data classification, second section is data clustering and third one is frequent pattern mining that each part will be studied separately. Taxonomy of data analyzingtechnoques by machine learning is as Figure 2.1.

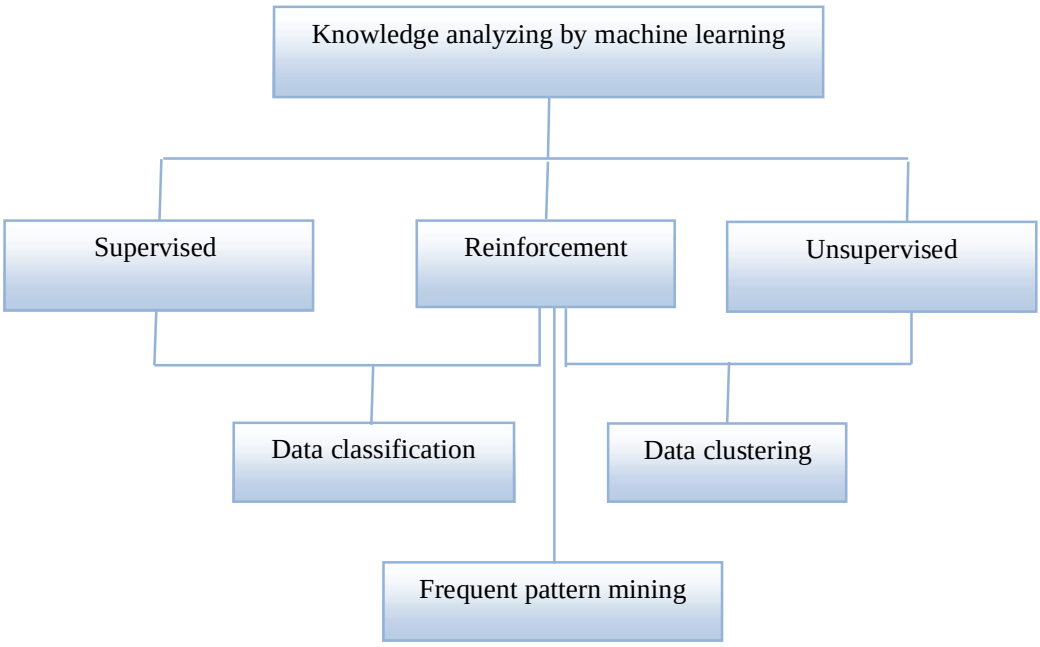

Figure 2.1 Taxonomy of data analyzing using machine learning

2.2 DATA CLASSIFICATION

Data classification is a process that can categorize data to achieve the relationship between attributes and extract the suitable rules for prediction process. Many

algorithms have been presented for data classifications which are referred to as follows:

2.2.1 Decision Tree

Decision tree is a tree which classifies samples on the basis of feature of each sample. Each node in decision tree indicates a feature in dataset which should be classified and each branch shows the value that the node can gain. The main problem of the decision tree is the high expense of its fabrication which is the NP-Complete.

Different methods have been created to find more suitable features which can be referred to Guyon and clisseff (2003). One of the most evident algorithms of decision tree fabrication is the C4.5. The last comparison which the decision tree has been made with other methods is given in Quinlan (1993). Elomaa (1999) used the binary discretization for C4.5 and obtained half training time with multi-splitting. Zheng improved another feature of decision trees; he improved classification accuracy of the model by creating at-least *M* of *N* features (Zheng 2000).Baik 2004 presented an agent-based approach for the distributed learning of decision trees and finally Yıldız (2007) study parallelization of the C4.5 algorithm. C4.5 decision tree has different application in other science such as effective network intrusion detection (Gandhi et al. 2010) and source code analysis (Taherkhani 2011).

2.2.2 Perception Method

Another method is perception method that perceptional meaning has been studied by (Rosenblatt 1962). First, a single layer perception is studied and then it will deal with a multilayer perception. If x_1 and x_2 are inputs and w_1 and w_2 are the weights (between [-1, 1], then, the output is $\Sigma x_1 w_1$ and output acts as a threshold and if the value of output is more than the threshold, output 1; otherwise is 0.

There are many works for optimization of this method such as Littlestone and Warmuth for weight updating (1994), Gardner and Dorling (1998). Freand and Schapir (1999) presented a new algorithm called the voted perception which saved

much information about training and they were used for predicting. This method should be repeated to give/provide a better answer and was good for binary models but should be practiced for higher classes and another work is perceptron-like linear algorithms for reaching the best accuracy by "anytime online algorithm" (Kivinen 2002).

One of the useful model of perception is Multilayer perception that it was presented for solving the above problem Freund and Schapire (1999). Multilayer perception was created by Rumelhart, Hinton and Williams (1986); and Wieland and Mischele (2008) combined this model with fuzzy technique to find a better fit of the model. In the next works, hidden layers were used for solving complex problems and one can calculate better and more precise answer by increasing and decreasing the number of neurons Kayri and Cokluk (2010).

2.2.3 Naïve Bayesian Classifier

Naïve Bayesian network (NB) is another method that is a simple network that uses direct acyclic graphs with only one parent and some children with strong independent hypothesis of children node inside other parts (Good 1950). Cestnik, Konenenlw and Braho (1987) used for machine learning at the first time. Advantage of this method is the training time of calculation. If features are not numerical, they should be preprocessed (Kotsiantis, Kanellopoulos and Pintelas (2006). However, the major form of this method is not consistent with datasets with high features, because it needs more space and time for creating of the related graphs. The major form need to preprocessing in most cases (Kotsiantis 2007). For another application that has been used from NB can refer to Bhasker, Kumud and Pardasani (2010). For studying about another method is referred to Archaux, Martin and Khenchaf (2004).

There are several applications that have used from CLA. These applications have represented for image processing and wireless sensor network classification, for example, Abin, Fotouhi and Kasaei (2008) used CLA for skin classify and segmentation, Meshkboo and Kangavari (2010) for video data mining and Mojaradi, Lucas and Varshosaz (2004) for post classification satellite imagery. Taxonomy of

data classification is as Figure 2.2 and Table 2.1 present some applications of data classification techniques.

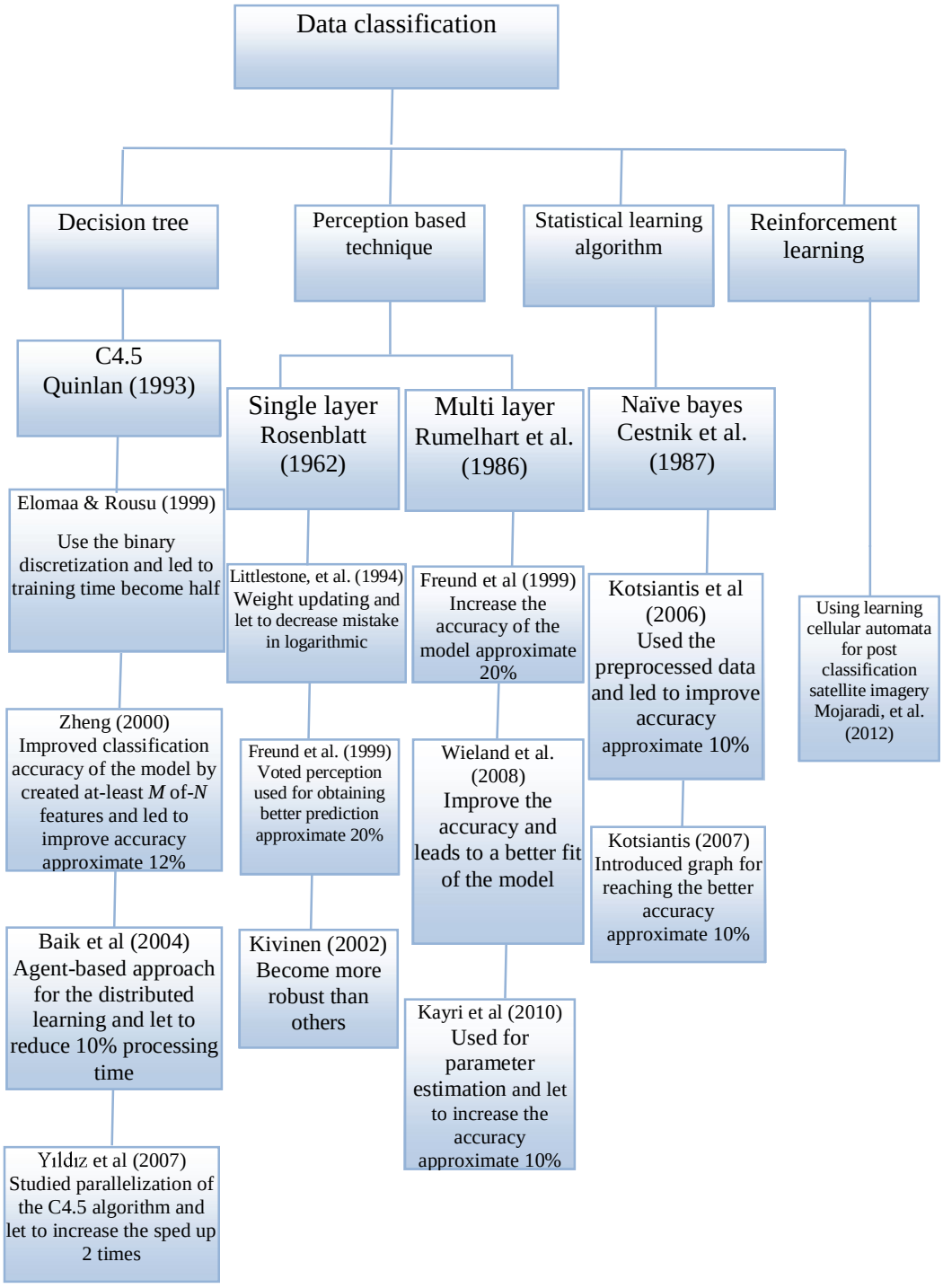

Figure 2.2 Taxonomy of the data classification

Table 2.1 Some applications of the data classification techniques

Publication year	The method	Domain	Category
Taherkhani (2011)	C4.5 decision tree	source code analysis	Use for classifier
Gandhi et al. (2010)	C4.5 decision tree	effective network intrusion detection	Classifiers decision
Tokumaru et al. (2009)	C4.5 decision tree	Product-impression analysis	Use for classifier
Gangrade, et al. (2009)	C4.5 decision tree	Privacy-preserving classification problem	Use for classifier
Atla et al. (2011)	Naïve Bayes	noise sensitivity	Noise classification
Xu et al. (2010)	Naïve Bayes	P2P traffic identification	Traffic classification
Gu et at. (2006)	Naïve Bayes	Proper smoothing for Information extraction	Information classification
Forman (2003)	Naïve Bayes	Feature selection metrics	Text classification
Chou et al. (2010)	Artificial neural network	Integrating web mining	Web cassification
Mitra et al. (2007)	Artificial neural network	Content-based audio classification	Audio classification
Mojaradi, et al. (2004)	Cellular learning automata	post classification satellite imagery	Image classification
Esnaashari et al. (2011)	Cellular learning automata	Mobile wireless sensor networks	Sensor networks classification
Meshkboo et al. (2010)	Cellular learning automata	Video data mining	Image classification
Beigy et al. (2009)	Cellular learning automata	Channel assignment algorithms	Channel classification
Motiee et al. (2008)	Cellular learning automata	Identification of web communities	Web classification

2.3 DATA CLUSTERING

Data clustering means a dataset is categorized/divided to separate cluster in terms of similarity or more precisely, the partitioning of a data set into subsets or clusters, so that the data in each subset, ideally, share some common trait. Clustering has different application in other science such as separation of medical images Mary and Kasmir (2010), urban development images clustering Samadzadegan, Saeedi and Hasani (2010), information retrieval Handl and Meyer (2002), and etc. Different methods for data clustering are present which perform clustering on the basis of type of data, form of cluster, data distance. Clustering is regarded as the unsupervised techniques in which multidimensional data is divided into some clusters on the basis of similarity or dissimilarity criteria Han and Yin (2001). Generally there are several methods for clustering as follow.

2.3.1 Hierarchical Method

In the hierarchical method, clustering has a single-link Sneath and Sokal (1973) and a complete-link (King 1967) structure, and the clustering is done by moving on its branches. Therefore, the complete-link algorithm produces tightly bound clusters (Baeza et al. 1992), while using a sparse graph that Guha, Rastogi and Shim (1999) based his work. However, it was Kurita who improved the single-link method from angle of the time complexity, (Kurita 1991), and Eddy used a large dataset for the single-link method (Eddy et al. 1994).

2.3.2 Partitioning Method

In the partitioning method, data is divided into some components on the basis of a similarity and put into clusters. An example of this is the simple method first used by James MacQueen called *k*-means (MacQueen 1967). There is another variation of the *k*-means algorithm that is called dynamic clustering algorithm by Diday (1973) and Symon (1977). After that, Pal, Bezdek and Tsao (1993) improved the K-mean algorithm by using mean data as the cluster center.

Finally, Han and Yin (2001) improved k-mean by the spatial clustering method, and eventually, Kumar introduced parallel K-mean clustering for improving the accuracy of the model. The valuation function was an error squares function and operates in two phases. In the first phase, each data belonging to the available clusters is studied thoroughly, and in the second phase, cluster centers are re-clustered with attention to the data of each cluster so that changes of the valuation function are less than the predefined value.

2.3.3 Ant Colony and Cellular Approach

Additionally, another method which has been scrutinized by the researchers is the ant colony and cellular method. In this clustering, data is put in mesh cells and clustered on the basis of their similarities (Wang et al., 1997). An example of this is in nature when, by using some other clustering methods, a combination of mesh structure and smart grouping has been observed, such as in the behavior of ants, movement of birds and fish, in which the speed of clustering depends on the close distance between the data. In this case, convergence to the answer may be slow.

Today, methods based on the behavior of ant cells, which are called the ASM (Ant Sleeping Model) algorithm, were cataloged by Chen, Xu, Chen and He (2004) as well as Moere and Clayden (2005) and Zhang, Peng and Zheng (2006) in a combination of cellular and ants methods. A method, based on the ant colony, was presented for the first time by Deneuborg, Goss Franks, Detrain and Chretian (1990) in which populations of ants that move randomly on the network are permitted to remove and add data items for data clustering. Based on this method, another model was developed by Lumer and Faieta (1994), resulting in a new algorithm called LF, which was presented and used for data analysis and search. In these altered methods, ants move on the network without data; therefore, there was a need for space for additional memory and more calculations. Another method was introduced by Kennedy and Eberhart (1995) called Particle Swarm Optimization (PSO). Using PSO, Shi (1998) simplified and optimized the algorithm. However before all of this, Nelder and Mead (1965) was one of the other methods that had been used for clustering. Kao,

Zahara and Kao (2008) combined K-means, Nelder–Mead and PSO to improve his latest method.

A new technique was introduced by (Xu, Chen and He (2005) which displayed another method, called A4C (Ant Clustering Embedded in Cellular automata). In A4C, an ant compares its own space to the other ants to find the best neighborhood to sleep. For this method, each data item is allocated to one ant and another are placed randomly in two–dimensional cells. Each ant moves on the network from cell to cell and has two active and resting states. The researcher calculates the probability of stagnancy and activity after the ant is placed in each new position on the basis of data proportion and similarity function. Each ant moves randomly on a cellular medium to find neighbors and to find similarities, there is no mechanism at first in the ants-based methods, which can be regarded as basis for targeted movement of ants in a cell net surface and cause to increase time of responding and delayed convergence to an answer in clustering. This time weakness can be frequently observed, where the number of clusters is high and similarity of data is low.

Some application of CLA for clustering is wireless sensor network Enami-Eraghi, Akbari-Torkestani, Meybodi, Fathy- Navid (2011) and Esnaashari & Meybodi (2008). Table 2.2 present some applications of the data clustering techniques as well as taxonomy of data clustering will be studied in Figure 2.3.

Table 2.2 Some applications of the data clustering techniques

Publication year	The method	Domain	Category
Samadzadegan et al. (2010)	Ant colony	Urban development	Images clustering
Mary et al. (2010)	Ant colony	Medical data	Images clustering
Handl et al. (2002)	Ant colony	Information retrieval	Information clustering
Chen et al. (1998)	K-means	Image segmentation	Images clustering
Liu et al. (2007)	K-means	Network traffic	Traffic clustering
Lima et al. (2010)	K-means	Network anomaly detection	Network anomaly clustering
Maarek et al. (2000)	Complete-link	Ephemeral document clustering	Web clustering
Hamou et al. (2010)	Cellular approach	Text mining	Text clustering
Enami-Eraghi et al. (2011)	Cellular learning automata	Channel Assignment Algorithms for Wireless Mobile Ad Hoc Networks	Wireless Mobile Ad Hoc Networks
Motiee et al. (2008)	Cellular learning automata	Identification of web communities	Web classification
Esnaashari, et al. (2008)	Cellular learning automata	Clustering Algorithm for Wireless Sensor Networks	Sensor network clustering

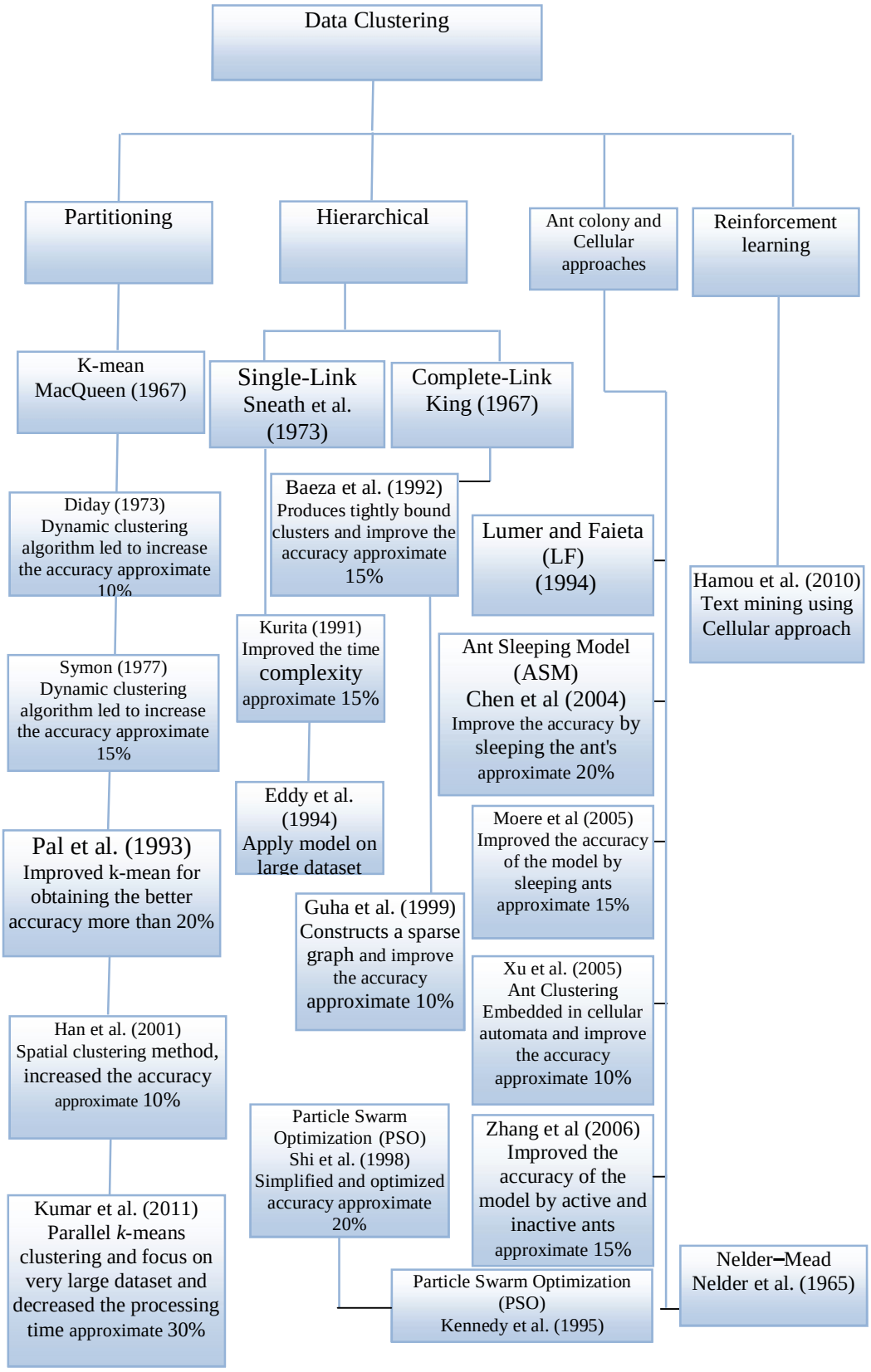

Figure 2.3 Taxonomy of the data clustering