Analysis of Various Clustering and Classification Algorithms in Datamining

Sandhia Valsala, Bindhya Thomas and Jissy Ann George

College of Computer Studies AMA International University Salmabad, Kingdom of Bahrain

Abstract:
Clustering and classification of data is a difficult problem that is related to various fields and applications. Challenge is greater, as input space dimensions become larger and feature scales are different from each other. The term "classification" is frequently used as an algorithm for all data mining tasks [1]. Instead, it is best to use the term to refer to the category of supervised learning algorithms used to search interesting data patterns. While classification algorithms have become very popular and ubiquitous in DM research, it is just but one of the many types of algorithms available to solve a specific type of DM task [12].
In this paper various clustering and classification algorithms are going to be addressed in detail. A detailed survey on existing algorithms will be made and the scalability of some of the existing classification algorithms will be examined.

Keywords
DM, clustering, classification, supervised learning, scalability

1. Introduction

Data classification comprises of several methods which mainly relies on the need for a particular application.How a method is selected varies on the amount of data and various classes present. These algorithms are aimed at serving a specific purpose.
Clustering of data relates to identifying similar data based on theirattribute. Numerous methods exist for clustering algorithms which can be used in a variety of applications [13]. The main difference between these algorithms lies in the two types of clustering algorithms, mainly, hierarchical and partitioning. The advantage of hierarchical clustering methods lies in its flexibility as the number of clusters is not needed as input. The disadvantage lies in its robustness as compared to partitioning algorithms. An error created in the initial steps continues into the later steps and eventually to the end of the program thus leading to a major error. Errors created in this way using data sets would result in misleading conclusions. Another disadvantage lies in the computational complexity rate, which is very large and thus cannot be used in real life situations.
When data is used to classify data, this is known as training data, which is used for automating data required for the generation of systems to classify data that will be used later on. This grouping relies on the likeness of incoming data to the training data [12]. The main challenge thus lies in creating systems which can group incoming data accurately.Classification and clustering are similar in most cases. The main difference lies in how the data are classified. Class boundaries need to be identified in classification whereas in clustering, the main task is the identification of the characteristics of the cluster. To achieve automatic classification systems, patterns need to be detected these patterns then needs to be studied as to how they are similar to significant classes, rather than just partitioning these data [17]. Many solutions have been proposed for both partitioning and clustering [7]. This paper makes an analysis of the various KDD Techniques.

2. Analysis of Existing KDD Techniques

DM and KDD fields are relatively new, different authors appear to survey methods in different ways [2]. Table 2.1 summarizes the various methods used for data mining.

Table 2.1Various methods used for data mining

AUTHORS	*METHODS FOR DATA MINING*
Fayyad	Predictive Modeling; Clustering; Summarization; Dependency Modeling; Change & Deviation Detection
Goebel &Gruenwald	Statistical Models; Case-Based Reasoning; Neural Networks; Decision Trees; Rule Induction; Bayesian Belief Networks; Genetic algorithms; Fuzzy Sets; Rough Sets
Aggarwal& Yu	Association rules, Clustering, Classification

In the following paper, Aggarwal& Yu's survey method was helpful in organizing the various methods. A survey of the various methods and techniques for data mining using Aggarwal& Yu's method is used in this paper to describe the various approaches, their strengths and weaknesses.

Manuscript received November 5, 2012
Manuscript revised November 20, 2012

Table 2.2 Analysis of the various KDD Methods

KDD Methods	***Advantages***	***Disadvantages***
Association Rules: In this, data correlations are found out between data items in different attributes. These represent the rules that associate between sets of items with some minimum level confidence factor. For example, an association rule is a statement that states: "90% [confidence factor] of transactions that purchase bread and butter [antecedent] also purchase milk [consequent]".	Generates very understandable data, and has become very popular in market analysis (also known as basket analysis).	Input and output is intensive for repeated passes over the database.
Clustering: In these kinds of algorithms, the data is groups' similar records into segments which are similar with respect to a group of points. Each of these segments may be treated differently. These algorithms are also known as unsupervised learning algorithms, and they rely heavily on statistical methods. Gray &Orlowska claim that clustering algorithms are typically in one of the following two forms: Hierarchical clustering: In this classification, a group of the data set is formed by either dividing or combining old clusters into new clusters. Partition-clustering: In this kind of clustering, the data set is divided into a particular number of clusters which is done by selecting and determining a partitioning based on the representative point for each cluster. The evaluation of this is carried out according to metric-distance- or model- based criteria. The best k-way partition is obtained using Metric-distance techniques so that data which is found in a block of the partition are closer in distance than to data which are found in other clusters. Model-based techniques find the best fit of a hypothesized model for each of the clusters, to each cluster. These are measured using probability measures which analyze the extent to which a model fits to a cluster.	Hierarchical Clustering: Hierarchical clustering generates clear rules from the data, without supervision or pre-defined labels. Partition-clustering: Can produce results that promise to be better than rule-based clustering methods.	Hierarchical Clustering: Partition clustering algorithms may outperform hierarchical rule-based algorithms. Partition-clustering: Like hierarchical clustering, finding rules that optimally covers all data clusters is an NP-hard problem.
Classification: Human supervision is required in these kinds of algorithmssuch that it can learn to classify data into certain defined classes. For example, when certain case histories of patients are given, the systems should be able to identify the best possible treatment for the patient. These can be further divided into: Decision Tree: These classifiers use a decision tree to partition data. This process is continued until each partition will have examples from a particular class. The split point in a decision tree is a non-leaf node in the tree that uses some condition (also known as predicates which represents the key factor in deciding how the data is divided. The terminal nodes in a decision tree contain tuples in the same class. k-Nearest Neighbor: The nearest neighbors are determined in this and a class label is assigned to the majority. This algorithm gives importance to the factor of locality. DNF Rules: On the left hand side of these rules, it specifies a union of a set of possible non-disjoint regions in the feature space; the right hand side corresponds to the class label. These algorithms aim to generate a minimal and complete rule set that is consistent with the training data. Neural networks: These networks are loosely modeled after the human brain. It is a data structure of functions that given one or more weighted inputs, produces an output which is a class label. Individual pieces of data are fed into the neural network, and the functions in the network are modified (learning) according to the error rates of the output produced. This approach usually results in long training times even if the data set is small.	Decision tree algorithms scale well, run fast, and produce highly interpretable results. k-Nearest Neighbor algorithms are easy to implement, and its results are interpretable. DNF Rules generate a small set of consistent rules, so algorithm scales well. Neural networks produce accurate results.	Decision tree algorithms suffer because they may find local maximas, producing inaccurate results. k-Nearest Neighbor algorithms produce huge models for small data sets. Scalability is a serious concern for these algorithms. Finding DNF Rules can be computationally expensive. Neural networks require long training times and produce hard to understand results. These algorithms are not scaleable, and significant data pre-processing is required.
Genetic algorithms: These algorithms are used to formulate hypotheses about dependencies between variables. In this type of algorithm numerous problem solutions compete with each other, the best among the lot are selected and then later joined together so that the solution set will improve over time. Bayesian networks: These are directed acyclic graphs that represent probability distributions, which can be used to represent expert knowledge. The nodes represent attribute variables and states (events), and edges represent probabilistic dependencies between them. Each node has local probability distributions, and arcs are drawn from cause	Genetic algorithms produce well tested accurate results. Bayesian networks produce easy to understand results. Fuzzy set models are less prone to local maximas than decision tree algorithms, and like rough set models, they cope with uncertainty better than any other algorithm.	Genetic algorithms arrive at results through evolutionary programming, and the results are often hard to understand. Must gather conditional probabilities from domain experts. Rough and Fuzzy sets require other algorithms to work.

nodes to effect nodes. Rough and Fuzzy Sets: These are mathematical concepts that deal with uncertainty. For Rough set models, an upper and lower bound is defined, each of which has members and non-members, respectively. The upper bound of a rough set is the union between the lower bound and the boundary region. Members of the boundary region are "possible members" of a set.		

Machine Learning

Learning paradigms

The learning paradigms can be classified into three based on the type of learning tasksinvolved [18].

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

1. Supervised learning

This kind of learning paradigms provides appropriate training patterns and also the outputs that are expected from the same. In this, the training set comprises of input patterns, the expected results.

In supervised learning the training set consists of input patterns as well as their correct results which are obtained in the form of the activation of all output neurons. Thus, for each training set that is fed into the network the output, can directly be compared with the correct solution and and the network weights can be changed. On minimizing the cost for the class of neural networks called Multi-Layer Perceptron's, the back propagation algorithm for training neural networks is obtained.

Uses of the supervised learning are mainly in pattern recognition, regression and for use in sequential data, such as speech and recognition of gestures.This kind of learning can be attributed to a teacher student learning form wherein the teacher provides feedback on the work or data presented by the student.

2. Unsupervised learning

This kind of learning paradigms provides input patterns to the network but without any feedback or help. This learning paradigm cannot be applied to all problems. With just the input patterns give, the network has to identify similar patterns and categorize them.

The training set comprises of input patterns, and the network tries to detect similarities and to generate classes.

3. Reinforcement learning

This kind of learning paradigms provides feedback to the network, if the system is functioning correctly or not. In reinforcement learning the network receives a logical or a real value after network completion of a sequence, which defines whether the result is right or wrong. Thus this procedure should be more effective than unsupervised learning since the network receives specific criteria for problem-solving.

The training set consists of input patterns, and on completing a sequence, a value is sent back to the network to indicate whether the result was right or wrong and, possibly, how right or wrong it was.ANNs are frequently used in reinforcement learning as part of the overall algorithm. This learning paradigm is mainly used in games, sequential decision making tasks and control problems.

Below ART, one of the types of unsupervised learning model will be analyzed in detail.

ART (Adaptive Resonance Theory) is a unique neural network paradigm. It is a self organizing network that can add learning without necessarily overwriting what is already learned. It clusters similar vectors and is capable of adding new clusters as necessary. ART I allows only binary vectors to be clustered. ART II is a (complicated) continuous version. The original description of ART by Grossberg is very complicated and is oriented towards the biological process of cognition, however we will study the simplified, problem solving version of ART (as is commonly used).

ARTconsistsofF1units(inputunits),F2units(clusterunits)a ndaweightmatrix (real)bwhichconnectsthemfrominputtocluster,andaweight matrix(alsobinary)t which connectsthem from cluster to input. The initial weights for b and tare set to a smallvalue and1 respectively. The b matrix is regarded as the long term memoryand the t matrixisregardedastheshortterm memory.

The steps of ART are vector input, "recognition" by the F2 units that identifies the winning cluster unit, "comparison" of the winning cluster with the input vector which decides through use of a "vigilance" parameter whether to assign the vector to this cluster or disable the cluster from competing for this vector, and finally weight changes to the winning cluster.

There are two learningalternatives. Thefirst is fast learning, where vigilance parameter= 1. Thesecondis slowlearning,wherevigilance parameter<1. Wewill usefastlearning.

There are a couple of easy tricks to improve ART 1 learning. You can sort the training vectors from most to least dense (most 1's to fewest 1's). Or, you can use complement coding which simply doubles the vector size and concatenates its

complement to it. This makes the number of 1's in each input vector the same (= n).

Types of ART

1. ART 1

The inputs to the networks are binary in the case of ART 1, which is also the simplest.[2].

2. ART 2

The inputs to this type of networks are continuous which in turn extends network capabilities. [3]

3. ART 2-A

A refined version of ART 2 is ART 2-A and this has a very good runtime and the results only slightly differ to the ART-2 implementation [4].

4. ART 3

ART 3 forms on ART-2 by simulating rudimentary neurotransmitter regulation of synaptic activity [5].

5. Fuzzy ART

A key improvement into ART's pattern recognition is obtained through generalization in Fuzzy ART which implements fuzzy logic. Another of its enhanced feature is complement coding which helps to prevent inefficient and unnecessary category proliferation [6] and is the manner in which the patterns are classified.

6. ARTMAP

A combination of ART-1 and ART-2 units has been used in the design of ARTMAP, which is also known as Predictive ART. This has led to the formation of a supervised learning structure where the input data is taken by the first unit and the correct output data is taken by the second unit which is then used to make the correct classification by making the minimum possible adjustment of the vigilance parameter in the first unit [7].

7. Fuzzy ARTMAP

ARTMAP implementation with the help of fuzzy ART units led to the design on Fuzzy ARTMAP. The implementation of the same has resulted in an increase in efficiency. [8].

The ART system makes use of an unsupervised learning paradigm. The main components of the ART system comprise of a vigilance parameter, reset module, comparison and recognition field of neurons. The key factor in the ART system is the vigilance parameter since a higher value of vigilance yields highly detailed memories (numerous, fine-grained categories), while a less vigilance value results in more general memories (less, more-general categories).

One of the parameters, namely, the comparison field makes use of an input vector (a one-dimensional array of values) and this is then sent to its best match to another parameter, which is the recognition field. A match to this would be the single neuron whose set of weights (weight vector) equals the input vector. A negative signal is emitted by each recognition field neuron to the other recognition field neurons which in turn results in the output being stopped. This is the manner in which the recognition field parameter exhibits lateral inhibition, which enables each neuron in it to represent a category to which input vectors are classified. The strength of the recognition match to the vigilance parameter is compared after the classification is done with the input vector. Only on meeting the threshold, will the training start else if the match with the vigilance parameter is not met; the firing recognition neuron is inhibited until a new input vector is applied. The commencement of training is met only on finishing the search.

It can be seen that, while searching, until a recognition match for the vigilance parameter is found, the neurons used for recognition are disabled one by one by the reset function. There are cases in which a committed recognition neuron's match does not meet the vigilance threshold, and then an uncommitted neuron isset into the committed state and adjusted such that it will match the input vector.

Conclusion

A detailed analysis of various clustering techniques and its advantages and disadvantages were conducted wherein the drawbacks of each technique was considered. A detailed look was taken into the various learning paradigms of which unsupervised learning paradigm were considered in detail. ART is a type of unsupervised learning technique in which various

A detailed look into unsupervised learning technique ART and its types were also conducted.

REFERENCES

[1] Holsheimer, M., Kersten, M., Mannila, H., Toivonen, H. "A Perspective on Databases and Data Mining", Proceedings KDD '95.

[2] Carpenter, G.A. &Grossberg, S. (2003), Adaptive Resonance Theory, In M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, Second Edition. Cambridge, MA: MIT Press

[3] Grossberg, S. (1987), Competitive learning: From interactive activation to adaptive resonance, Cognitive Science (Publication).

[4] Carpenter, G.A. &Grossberg, S. (1987), ART 2: Self-organization of stable category recognition codes for analog input patterns, Applied Optics.

[5] Carpenter, G.A., Grossberg, S., & Rosen, D.B. (1991), ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition, Neural Networks (Publication).

[6] Carpenter, G.A. &Grossberg, S. (1990), ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures, Neural Networks (Publication)
[7] Carpenter, G.A., Grossberg, S., & Rosen, D.B. (1991b), Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system, Neural Networks (Publication)
[8] Carpenter, G.A., Grossberg, S., & Reynolds, J.H. (1991), ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network, Neural Networks (Publication)
[9] Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., & Rosen, D.B. (1992), Fuzzy ARTMAP: A neural network architecture.

Web Pages
ftp://ftp.fas.sfu.ca/pub/cs/han/kdd
http://www.data-miner.com/
http://www.cs.purdue.edu/homes/ayg/CS590D/resources.html
http://www.kd1.com/

Text Books
Pieter Adriaans, DolfZantinge - Data Mining – Addison Wesley Longman – 1999
Arun K Pujari - Data mining techniques - Universities Press – 2001
Michael J. A. Berry, Gordon S. Linoff – Mastering Data Mining – Wiley Computer Publishing – 2001
Rhonda Delmater, Monte Hancock – Data Mining Explained – Butterworth-Heinemann – 2001
D. Kriesel – A Brief Introduction to Neural Networks

Other References
http://www.revue-texto.net/Reperes/Biblios/Forest_Biblio.html
http://cns-web.bu.edu/~steve/
http://www-users.cs.umn.edu/~desikan/research/dataminingoverview.html

Ms.Sandhia Valsalais presently associated with AMA International University, Bahrain as the Department Head of Computer Science.She holds a Master's degree in Computer Applications from Bharatiyar University,Coimbatore and is currently pursuing her PhD from Karpagam UniversityCoimbatore.

Ms. Bindhya Thomas is presently associated with AMA International University, Bahrain as Lecturer in the Computer Science Department. She holds a Master's degree in Computer Science from Bharatiyar University,Coimbatoreand is currently pursuing her PhD from Karpagam UniversityCoimbatore.

Ms. Jissy Ann George is presently associated with AMA International University, Bahrain as Lecturer in the Computer Science Department. She holds a Master's degree in ITfrom AlagappaUniversity and is currently pursuing her PhD from Karpagam UniversityCoimbatore.