Algorithms for Data Mining: A Survey

Shruti A. Vagare 1,Mangesh M. Ghonge2

1Department of Computer science & Engineering, JCOET Yavatmal, India, vagareshruti@gmail.com,
2Professor, Department of Computer Engineering, JCOET Yavatmal, India, mangesh.cse@gmail.com,

Abstract- There are many algorithms for mining data that are being constantly developed and improved by research communities and industrial organizations worldwide. This paper surveys the most used algorithms for data mining in order to point out their similarities and differences, their advantages and drawbacks, and studying for each algorithm.

Index Term -Data mining; Association rules; Clustering; Classification

1. **INTRODUCTION**

The increasing improvement of computer technologies and the introduction of the World Wide Web have led to a massive growth of data and information, to a point where traditional methods for data exploring and analyzing have become highly consuming in terms of time and resources, and even ineffective in many cases, especially with high dimensional databases such as astronomical databases.[1] Which created a huge need for data mining technologies to search for valuable knowledge in large volumes of data.There are many data mining techniques and algorithms, but choosing the right technique for the right problem is a necessary step that is based on defining these parameters: the type of data, the type of the database, and the goals behind applying the data mining model. The main emphasis that influence most the knowledge discovery process.

According to the oxford dictionary, data refer to the facts and statistics collected for reference or analysis, in data mining, these facts and statistics are represented by a collection of objects described by their attributes. In the literature, an object is also known as: record, point, entity, or instance. The nature of the data have a huge impact on choosing a particular data mining task over another, for example some classifiers that require the definition of distance measure between objects perform well on numerical data.[10]

Data mining is defined as a step in knowledge discovery in databases (KDD) that explore databases in order to reveal hidden information or predict behavior based on frequent patterns.(fig 1.)[9]

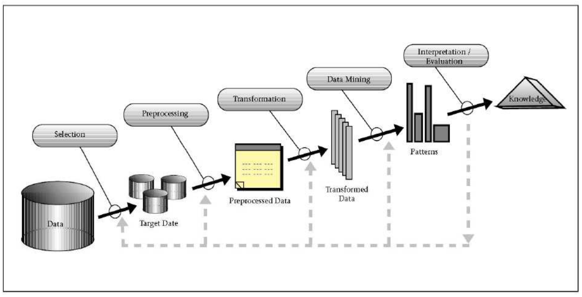


Fig 1: the process of Knowledge Discovery in Databases[9]

A preprocessing step to improve data quality must be performed to help obtaining better results, this step include removing noise and outliers from data, and dealing with missing values and duplicate records, because incoherent and incomplete data can lead obviously to extracting unreliable knowledge. Some preprocessing data techniques include:

•**Aggregation**: combining two or more attributes or objects into a single one.

•**Sampling**: using a representative sample instead of the whole datasets.

•**Dimensionality reduction**: eliminate irrelevant features or reduce noise to optimize performance and facilitate data visualization.

1.DATA MINING:

Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data. Data mining products can be very powerful tools; they are not self sufficient applications. To be successful, data mining requires analytical specialists and skilled technical that can structure the analysis and interpret the output that is created. Data mining is discovering the methods and patterns in large databases to guide decisions about future activities.[7]

The data used for data mining are usually organized in databases or datasets, there are different types of databases:

- **Relational**: objects have the same fixed attributes, and are connected.
- **Document**: objects are “term” vectors, Where each term is a component of the vector and its value is determined by the number of occurrences in the document
- **Transactional**: objects or transactions are a set of items, such as the set of products purchased by a customer at a grocery store.
- **Spatial and/or temporal databases**: Each type of these has a set of models and algorithmsbthat are best suited for mining its data.

2. ALGORITHMS:

Here there are several data mining tasks that are used for either predictive or descriptive purposes, in this section, we present three of them which are: classification (predictive), clustering (predictive), and association rule analysis (descriptive).[10]

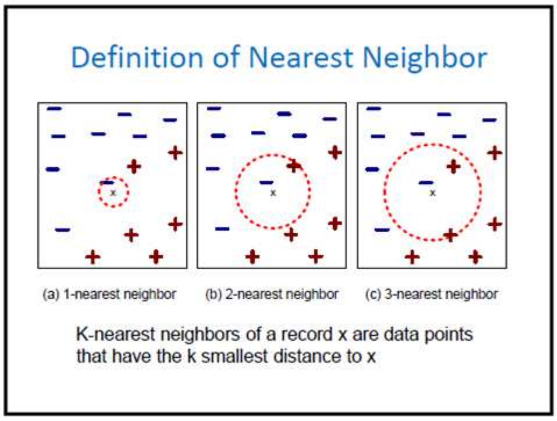
2.1 CLASSIFICATION:

Assigning an object to a certain class based on its similarity to previous examples of other objects. It can be done with reference to original data or based on a model of that data. Classification is one of the most widely used models in data Mining, Given a collection of records, each record defined by a set of attributes, and one of them represents the class, a class represents a homogeneous group of cases or records. The goal of classification is finding a model for the class attribute, as a function of the values of other attributes, so that a previously unseen record could be easily assigned to a class as accurately as possible.[3]

Suppose that an object is sampled with a set of different attributes, but the group to which the object belongs is unknown. Assuming its group can be determined from its attributes different algorithms can be used for classification process. A nearest neighbor classifier is a technique for classifying elements based on the classification of the elements.With the k-nearest neighbor technique, this is done by evaluating the k number of closest neighbors.The k-nearest neighbor classification algorithm can be expressed as:

```
k -number of nearest neighbors for each object
X -in the test set
do
calculate the distance D( X ,Y ) between X and every
object Y in the training set neighborhood
i.e  the k  neighbors in the training set closest to X
X.class = SelectClass(neighborhood )
end
```

The k-nearest neighbors’ algorithm is amongest the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. k is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose k to be an odd number as this avoids tied votes. Defination of k-nearest neighbors as shown in fig.2.


Definition of Nearest Neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

Fig.2 Defination of nearest Neighbor

2.1.1. ADVANTAGES:

1] To weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

2] The neighbors are taken from a set of objects for which the correct classification is known.

These classification algorithms can be implemented on different types of data sets like data of patients, [3]financial data according to performances. On the basis of the performance of these algorithms, these algorithms can also be used to detect the natural disasters like cloud bursting, earth quake,etc.

2.2CLUSTERING:

Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships.The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the “better” or more distinct the clustering.Clustering is the task of finding groups of similar or related objects, such that two objects from the same group are similar or related to one another and different from the objects in other groups.

Clustering objects and things into different groups is a common way to describe the world, here are some domains that it is widely used in[10]:

- **Biology:** biologists have always tried organizing all living things into groups based on their similarities of structure, origin…etc.

- **Information Retrieval:** the web contains billions of web pages, so that a query in a
search engine can returns millions of results, these results can be clustered into groups which can make it easy for the user to explore the results of his query.
- **Climate:** finding patterns in the atmosphere and ocean help understanding.
- **Psychology and medicine:** clustering can be used to determine different types of a disease,
and it can also be used for finding patterns in its temporal and spatial distribution.
- **Business:** segment customers into groups for additional analysis and target marketing.

Basic Algorithm of clustering:

The K-means clustering technique is very simple and we immediately begin with
a description of the basic algorithm.[3]The following are the steps :

Basic K-means Algorithm for finding *K* clusters.

1. Select *K* points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change

.

In the absence of numerical problems, this procedure always converges to a solution, although the solution is typically a local minimum. The following diagram gives an example of this. Figure a shows the case when the cluster centers coincide with the circle centers. This is a global minimum. Figure b shows a local minima.

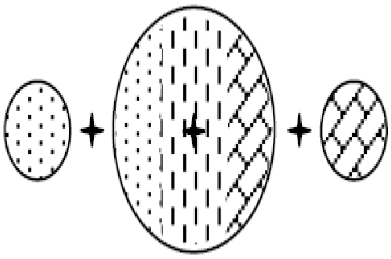


Fig a. A globally minimal clustering solution

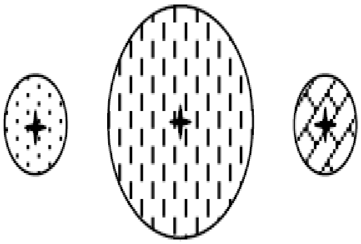


Fig b. A locally minimal clustering solution

2.2.1. ADVANTAGES:

1] K-means produces high accuracy.
2] Less computation time.

2.2.2 LIMITATIONS:

1] K-means has problems when clusters are of differing in Sizes, Densities, Non-globular shapes.
2] K-means has problems when the data contains Outliers.

2.3 ASSOCIATION RULE:

Association rule mining represents a data mining technique and its goal is to find interesting association or correlation relationships among a large set of data items.One of the most common data mining approaches is finding frequent item-sets in transactional databases, and deduct their corresponding association rules. An itemset is a collection of one or more items, a k-itemset is an itemset that contains k items.[4]

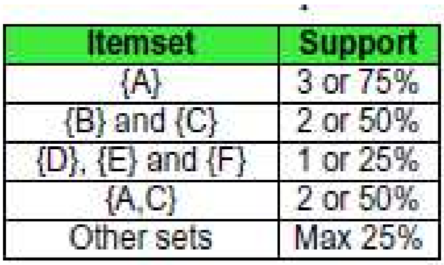
□

Association rule mining is a two-step process:

1. Find the frequent itemsets, i.e., the sets of items that have at least the minimum support s.
2. Use the frequent itemsets to generate (strong) association rules that satisfy the minimum support s and minimum confidence c.

Example of association rule:

1] Itemsets and their frequencies/supports:



Itemset	Support
{A}	3 or 75%
{B} and {C}	2 or 50%
{D}, {E} and {F}	1 or 25%
{A,C}	2 or 50%
Other sets	Max 25%

2] If the minimum support s and the minimum confidence c are both 50%, then we get association rules

A-> *C* [50%, 66.6%]
C->*A* [50%, 100%]

One of the widely used association rule discovery algorithms is the APRIORI algorithm, which uses candidate generation to find frequent itemsets, it considers that if an itemset is not frequent all of its superset is not frequent, which prevent monotonicity in the level-wise search of frequent itemsets.

Apriori algorithm:

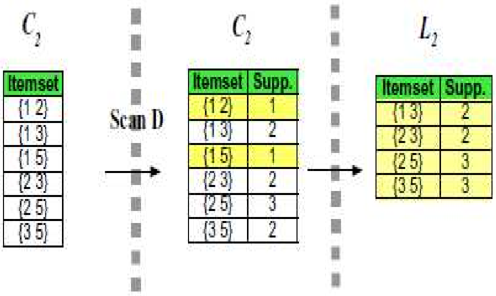
1]a basic algorithm for finding frequent itemsets for boolean association rules
2]based on levelwise search:

-iteratively find frequent itemsets with size from 1 to *k* (*k*-itemset)

Algorithm as follows:

```
Ck: Candidate itemsets of size k;
Lk : Frequent itemsets of size k
L1 = {frequent items};
for (k = 1; Lk !=∅; k++) do begin
Ck+1 = {candidates generated from Lk };
for each transaction t in the database do
increment the count of all candidates in Ck+1
that are contained in t;
Lk+1 = {candidates in Ck+1 with σ}
end
return ∪k Lk;
```

Example of Apriori :



Itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Itemset	Supp.
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

Itemset	Supp.
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	3

ADVANTAGES:

1] Express how items or objects are related to each other, and how they tend to group together.
2] Simple to understand.
3] Provide useful information.
4] Efficient discovery algorithms exist.

APPLICATIONS:

Market basket data analysis, cross-marketing, catalog design, loss-leader analysis,etc.

FUTURE WORK:

Due to the enormous success of various application areas of data mining, the field of data mining has been establishing itself as the major discipline of computer science and has shown interest potential for the future developments.

CONCLUSION:

In this paper we survey the various data mining algorithms. To perform an exhaustive survey of the literature in this domain is a very hard task that is both time and energy consuming, so the purpose of this paper was to briefly present some of the most noted algorithms in both industrial and research communities. These algorithms were have different domains of application and use different approaches and techniques, they remain very simple and efficient algorithms, and are being constantly improved by the research community. This review would be helpful to researchers to focus on the various issues of data mining. In future course, we will review the various classification algorithms and significance of evolutionary computing (genetic programming) approach in designing of efficient classification algorithms for data mining.

REFERENCES:

[1] What is R? Retrieved March 12, 2014 from www.rproject.org

[2] 2013 Data Miner Survey, Retrieved from March 11, 2014 from http://www.rexeranalytics.com/DataMinerSurvey-Results-2013.html

[3]Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao,Data mining techniques and applications – A decade review from 2000 to 2011, Elsevier Expert Systems with Applications, vol. 39, pp. 11303–11311, September 2012

[4] Wasilewska, A. - *APRIORI Algorithm*, Lecture Notes, http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf, accessed 10.01.2007

[5] Xiaojun Chen, Yunming Ye, Graham Williams,and Xiaofei Xu, A Survey of Open Source Data MiningSystems, PAKDD 2007 Workshops, LNAI 4819, pp. 3–14, 2007.

[6] Jiawei Han from Simon Fraser University and Fosca Gianotti and Dino Pedreschi
from University of Pisa,spring 2005

[7] Kantardzic, Mehmed ―Data Mining: Concepts, Models, Methods,
and Algorithms‖, John Wiley & Sons, 2003.

[8] Rakesh Agrawal, and Ramakrishnan Srikant, (1997), *Fast Algorithms for
Mining Association Rules*, In Proceedings of the 20th VLDB Conference,
pages 487-499, Santiago, Chile. http://www.almaden.ibm.com/cs/people/ragrawal/

[9] U. Fayyad, Piatetsky-Shapiro, and P. Smyth, From Data Mining to Knowledge Discovery in Databases, AI Magazine, vol. 17, pp. 37-54, 1996.

[10]P. Tan, M. Steinbach, V. Kumar. Introduction to Data Mining [PDF document]. Retrieved from lectures notes online web site:http://www.users.cs.umn.edu/~kumar/dmbook/index.php.