Various Strategies and Technical Aspects of Data Mining: A Theoretical Approach

Ekbal Rashid, Srikanta Patnaik and Arshad Usmani

Abstract In this paper, we are going to look at a very interesting aspect of database management, namely data mining and knowledge discovery. This area is attracting interest from not only researchers but also from the commercial world. The utility of data mining in commerce is more interesting than perhaps research areas. This has also raised many debates such as rights of privacy, legality and ethics, and rights to non-disclosure of information. It has someway opened a Pandora's box. Only, time will tell whether it is on the whole destructive or constructive; nonetheless, technology is not as such absolutely constructive or destructive; it only depends on how it is brought into use. In this paper, we have discussed about the technical aspects of data mining and what are the different strategies of data mining. Its sections give many technical aspects for various data mining and knowledge discovery methods, and we have given a rich array of examples and some are drawn from real-life applications.

Keywords Data mining · Classification · Algorithm · Clustering · Association

1 Introduction

Let us brief out first as to why is data mining necessary.

If one needs to talk to a manager or some Internet search on why data mining is necessary, there would be a variety of answers. Some would say it is very good for wealth generation, some would say it helps me to understand how market is

E. Rashid (✉) · A. Usmani
Cambridge Institute of Technology, Ranchi, India
e-mail: ekbalrashid2004@yahoo.com

A. Usmani
e-mail: ausmani@yahoo.co.in

S. Patnaik
SOA University, Bhubaneswar, Orissa, India
e-mail: patnaik_srikanta@yahoo.co.in

I.K. Sethi (ed.), *Computational Vision and Robotics*, Advances in Intelligent
Systems and Computing 332, DOI 10.1007/978-81-322-2196-8_6

behaving, and some would say it would be great for security purposes. That is to identify abnormal activities in the network. This is an important area of data mining in these days. In short, there is no doubt about the growing influence of this aspect of data management.

The rest of the paper is structured as follows: Sect. 1 gives introduction, Sect. 2 describes the type of data. In Sect. 3 we present data mining and statistical inferences. In Sect. 4, association and item sets discussed. Section 5 presents classification and clustering with conclusion.

1.1 What Is Data Mining then?

It is the generic term used to look for hidden patterns and trends in data that is not apparent in just summarizing the data [1]. We can look at patterns of students' performance in certain subjects. Whether students performing well in a particular subject will do good in some other one. Putting in a very broad sense, data mining is controlled by something called 'interestingness criteria.' Finding something or finding everything according which one does not know about according to some criteria. We start with a database. We give some interestingness criteria and then discover trends. The output should be one or more patterns that we did not know to exist. In data mining, we do not talk about any data mining query. Rather, it is the data mining algorithm that should give us something that we do not know. When we talk about patterns, we say something is a pattern and something is not a pattern. For this, we have to see what kind of data we are looking at. We also have to keep in mind what is the type of interestingness criteria that we are looking at.

2 Types of Data

The different types of data that we encounter are [2]:

Tabular or relational database in the form of tables. This is the most common form of data. Then, there are special data, having coordinates and attributes. There are temporal data which are data with a time tag associated with it, like the traffic in a network or database activity logs. Then, tree data like xml databases, sequence data such as information about genes. There are text and multimedia data.

When we talk about interestingness criteria, we can talk about frequency, that is how often something happens, or rarity which may point toward abnormal behavior. Co-relations between data, consistency, can also be considered to be interestingness criteria. A customer comes once every month is more consistent than a customer who comes ten times first month and not even a single time the second month and so on. Periodicity may be another interestingness criterion.

3 Data Mining and Statistical Inferences

When we talk about data mining, there is sometimes a confusion that data mining is the same as statistical inferences. There is fundamental difference between the two. In statistical inference, we start with a null hypothesis; that is, we make a hypothesis about the system, like if exams are in March, then the turnout will be higher. Based on this, we do sampling and this is very important in statistical inference. Based on this sampling, we verify or refute the hypothesis. There is some questionnaire that knows what to look for.

In data mining, we just have a huge data set. We do not have any null hypothesis. We use some interestingness criteria to mine this data set. Usually, there is no sampling done. The data set is mined at least once to look for patterns. We present a data mining algorithm for the data set. Sometimes it is also called ‘hypothesis discovery.’

4 Association and Item Sets

Two fundamental concepts which are of interest in data mining are of association and item sets [2] .We also have the concept of support and confidence. The support for a given rule R is the ratio of the number of occurrences of R given all occurrences of all rules. The confidence of a rule $x \rightarrow y$ is the ratio of the number of occurrences of y given x among all other occurrences given x. The apriori algorithm is used in data mining for frequent item sets. The apriori algorithm goes as such:

```
Given is a minimum support s as interestingness criteria.

Step 1. Search for all individual elements (1 element item set) that have
   a minimum support of s.
Step 2. REPEAT
a) From the results of i element item sets, search for the results of i + 1
   element item sets that have a support of s.
b) This becomes the set of all frequent (i + 1) item sets that are
   interesting.

Step 3. UNTIL item set size reaches maximum.
```

The property of item sets is that you consider item sets as atomic; that is, there is no point of ordering in the item sets. As for example, it does not matter whether a customer buys item 1 first or item 2 first, as long as they are bought together. From this, we can draw the inference that the two items are quite likely to be brought together in one piece. How can this information be useful? Say if one has a supermarket and data mining has resulted in the understanding that two particular items are likely to be bought together in a single packet, such as bags and books are likely to be bought together, then from this, it would be wise for the supermarket owner to place bags and books side by side in the supermarket so that it would be easier for people to select both of them.

However, if we are looking toward association rules then there is a sense of direction in it. If we say if X then Y, it is different from saying if Y then X. Hence, association mining requires two different thresholds, the minimum support as in the item set and the minimum confidence with which we can determine that a given association rule is interesting. The following is the method to mine association rules using apriori.

1. First use apriori to generate different item sets of different sizes.
2. At each iteration, divide the item set into two parts the LHS and the RHS. This gives us a rule of the form LHS → RHS.
3. The confidence of this rule is that support of LHS → RHS divided by the support of LHS.
4. The confidence of all those rules are discarded which have a support of less than that of minimum confidence.

5 Classification and Clustering

Next, there are the tasks of discovering a classification or of discovering clusters within a data set. So come the concepts of classification and clustering. Intuitively, they seem to do the same thing. However, there is a marked difference between the two. Classification maps elements to one of the sets of predetermined classes based on the difference among data elements belonging to different classes [3].

Clustering groups data elements into different groups based on the similarity between elements within a single group [4]. In classification, we know the classes beforehand while mostly in clustering we do not know how many clusters we are going to get. In data mining, we are interested in discovering classification. For example, suppose we have data about cricket matches that have been played in a particular city. Now, this city is notorious for its frequent changes in weather. Suppose we have data such as if the day is sunny and the temperature is 30°, play is continued. If the day is overcast and temperature is 17°, then play is abandoned; if day is sunny and temperature is 20°, then still the play is continued, so on and so forth. Now, the classification problem is that whether one can classify the weather conditions, that is, how the day was and the temperature into one of the two classification criteria, whether play will be organized or it will be abandoned.

For classification problem, we use the Hunt's algorithm for decision tree identification. It goes as follows:

```
Given N different element types and m different decision classes,
Step 1. For i ← 1 to n do
a. Add element i to the i-1 th element item sets from the previous
   iteration.
b. Identify the set of decision classes for each item set.
c. If an item set has only one decision class, then it is done, remove that
   item set from the subsequent iteration.
```

Now, we look into the methods of clustering. Clustering is philosophically different from classification. Classification is the method of increasing the differences between the elements so as to make them belong to different classes, while clustering is the process of increasing the similarities between different elements so as to form them into different clusters. So clustering essentially partitions data sets into several clusters. What is the property of a particular cluster? The property is that the similarity of different elements in one cluster is much more than the similarity between different elements across clusters. So members belonging to the same cluster are much more similar to one another than they are to some members of some other cluster. And there are several measures of similarities and most of which are reduced to geometric similarities by projecting these data sets into hypercubes or n-dimensional spaces and then using some kind of distance measures such as Euclidean distance measures or Manhattan distance to compute the similarity. The first kind of clustering algorithm is called the nearest neighbor clustering algorithm. This clustering algorithm takes a parameter called threshold or the maximum distance t between the members of a given cluster [5].

```
Given n elements x1, x2,… xn and given a threshold t,

Step 1. j ← 1, k ← 1, cluster = []
Step 2. Repeat
find the nearest neighbor of xj
let the nearest neighbor be in some cluster m
if distance to nearest neighbor is greater than t, then create a new cluster and
increment the number of clusters or assign it to the cluster m
j ← j-1
Step 3. until j > n
```

There is another kind of clustering technique which is also popular and it is called iterative conditional clustering. This differs from the nearest neighbor technique in the sense that here, the number of clusters are fixed beforehand. In the nearest neighbor technique, the numbers of clusters are not fixed beforehand. That means one does not know how many clusters one is going to get given a particular threshold and a data set.

Given n different elements and *k* different clusters and each with a center, this center means the centroid in the statistical sense.

Step 1. Assign each element to the closest cluster center
Step 2. After all assignments have been made, compute the cluster centoids for each of the cluster. This means one has to compute the average of all the points that have made up this cluster. Possibly, this will shift the centroid to a different location.
Step 3. Repeat the above two steps with the new centroids until the algorithm converges or stabilizes so that the centroids will stop shifting and then we know that we have found the best centroids for each of the clusters.

Iterative conditional clustering is essentially a technique where something like saying that suppose I have a data set, and suppose I want to create ten different clusters out of this data set, where would these clusters lie. On the other hand, nearest neighbor clustering technique would say suppose I have this data set and suppose I have some maximum distance, between elements that can lie between a data set, then how many clusters will I find. This concludes about technical aspects and various strategies of data mining and clustering.

References

1. Dunhum, M.H.,: Data Mining: Introductory and Advanced Topics, 3rd Impression. Pearson Education, India (2008)
2. Srinivasa, S.: Data Mining, Data Warehousing and Knowledge Discovery Basic Algorithms and Concepts. www.anilsoft.com
3. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2nd edn. Morgan Kaufmann Publishers, Massachusetts (2006)
4. Triantaphyllou, E.: Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques. Springer, Berlin (2009). ISBN:1441941738 9781441941732
5. Catal, C., Sevim, U., Diri, B.: Clustering and metrics threshold based software fault prediction of Unlabeled Program Modules. In: 6th International Conference on Information Technology: New generations, pp. 199–204. IEEE (2009)