National Conference on "Advanced Technologies in Computing and Networking"-ATCON-2015
Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

Data Mining Technique A Critical Survey (Knowledge Discovery from Data)

Prof. B. P. Deshmukh **Prof. Amrapal D. Chavan**

***Abstract* — Data mining technique can be defined as it is the activity to extract the some information from the large data base the aim is to discover the hidden information or actual relationship of the data to the metadata.**

Today we find application in wide range of business, medical technology, scientific and engineering application example in a case of medical utilization this database can be generated the typical pattern so it can help to determine the dieses .it is also help to determine or analysis the financial data.

***Key Words* — Data Mining, Naive bayes algorithm, the k-nearest neighbor algorithm, the Apriori algorithm.**

I. INTRODUCTION

Why Data Mining?

- Data collection and data availability
- Automated data collection tools, database systems, Web, computerized society.
- Large sources of big data
- Business:Web- e-commerce, transactions of finance
- Science: Remote sensing, bioinformatics, scientific application
- Commercial –Banking, loan, customer large data etc.

Data Mining is an analytic process designed to explore data (usually "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.

The process of data mining consists of three steps, **1.Exploration**

In this task the data preparation should be done by data cleaning data, data transformations, selecting of sub records but some time data should be large in size with their variables and attribute and performing some preliminary selection operations to bring the number of variables to a manageable range then depending on the nature of the problem this is the first stage of the process of data mining

2.Model building and validation

In this stage considering various models and finding the best one solution based on their predictive performance. This task look like the general application on the other hand reality is it take or involves a very elaborate process. There are a multiple techniques are used to achieve that goal. These techniques - which are often considered the starting approach to predictive data mining

3.Deployment

The final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

The concept of Data Mining is becoming increasingly popular. It covers the activities to organize knowledge gained through data mining models and present it in a way users can use it within decision making.[6]

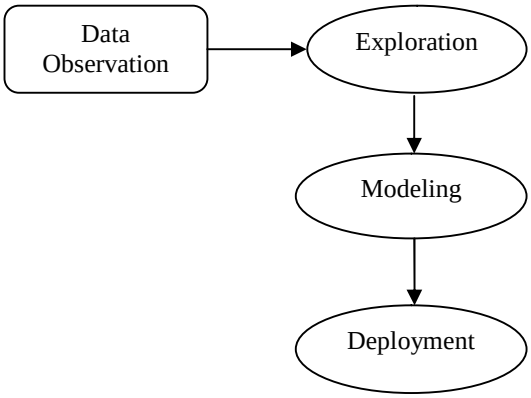

Fig. 1. Data mining operation

II. ALGORITHM

Data mining models and algorithm

I)THE K-NEAREST NEIGHBOR ALGORITHM

The nearest neighbor algorithm is a prediction technique which is just like forecast the prediction of weather which is nearest to our object record in the data base it is more easy to understand because it work same that of way of people thinking. so that we can state that nearest neighbor algorithm the object which are having the nearest location to each other having the similar prediction value [1]

The nearest neighbor algorithm. It is technique for identifying the element base on the classification of the element. the k-nearest neighbor technique, this is done by evaluating the k number of closest neighbors.

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 ← the k neighbors in the training set closest to X
X.class ← SelectClass(neighborhood)
End for [3]

II) NAIVE BAYES

Given a set of objects, each of which belongs to a known class, and each of which has a known vector of variables,
our aim is to construct a rule which will allow us to assign future objects to a class, given only the vectors of
Variables describing the future objects. Problems of this kind, called problems of supervised classification, are
ubiquitous, and many methods for constructing such rules have been developed. One very important one is the naive
Bayes method—also called idiot's Bayes simple Bayes, and independence Bayes. This method is important for
several reasons. It is very easy to construct, not needing any complicated iterative parameter estimation schemes.
This means it may be readily applied to huge data sets. It is easy to interpret, so users unskilled in classifier
technology can understand why it is making the classification it makes. And finally, it often does surprisingly well it may not
Probabilistic approaches to classification typically involve modeling the conditional probability
distribution P(C|D), where C ranges over classes and D over descriptions, in some language, of objects to be
classified. Given a description d of a particular object, we assign the class argmaxcP(C = c|D = d). A Bayesian
approach splits this posterior distribution into a prior distribution P(C) and a likelihood P(D|C):

$$\text{argmaxc } P(C = c|D = d) = \text{argmaxc } \frac{P(D = d|C = c)P(C = c)}{P(D = d)} \quad (1)$$

The denominator P(D = d) is a normalizing factor that can be ignored when determining the maximum a posteriori
class, as it does not depend on the class. The key term in Equation (1) is P(D = d|C = c), the likelihood of the given
description given the class (often abbreviated to P(d|c)). A Bayesian classifier estimates these likelihoods from
training data, but this typically requires some additional simplifying assumptions. For instance, in an attribute-value
representation (also called propositional or single-table representation), the individual is described by a vector of
values a1, . . . , an for a fixed set of attributes A1, . . . ,An. Determining P(D = d|C = c) here requires an estimate of
the joint probability P(A1 = a1, . . . ,An = an|C = c), abbreviated to P(a1, . . . ,an|c). This joint probability
distribution is problematic for two reasons:
(1) its size is exponential in the number of attributes n, and (2) it requires a complete training set, with several
examples for each possible description. These problems vanish if we can assume that all attributes are independent
given the class:

$$P(A1 = a1, \ldots, An = an|C = c) = \prod_{i=1}^{n} P(Ai = ai|C = c) \quad (2)$$

This assumption is usually called the naive Bayes assumption, and a Bayesian classifier using this assumption is called the naive Bayesian classifier, often abbreviated to 'naive Bayes'. Effectively, it means that we are ignoring
interactions between attributes within individuals of the same class.

III) THE APRIORI ALGORITHM

The data mining approach is to finding frequent itemset from transaction of dataset .if the frequent itemset are obtained which is straight forward to generate specified minimum data record Apriori is a seminal algorithm for finding frequent itemsets using candidate generation It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. Let the set of frequent itemsets of size k be Fk and their candidates be Ck Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets.

1. Generate Ck+1, candidates of frequent itemsets of size k +1, from the frequent itemsets of size k.
2. Scan the database and calculate the support of each candidate of frequent itemsets.
3. Add those itemsets that satisfies the minimum support requirement to Fk+1.

Apriori algorithm is given in the below figuer1:

F_1=(Frequent itemsets of cardinality 1);
for($k = 1; F_k \neq \phi; k++$) do begin
  C_{k+1} = apriori-gen(F_k); //New candidates
  for all transactions $t \in$ Database do begin
    C'_t = subset(C_{k+1}, t); //Candidates contained in t
    for all candidate $c \in C'_t$ do
      $c.count++$;
    end
    $F_{k+1} = \{C \in C_{k+1} \mid c.count \geq$ minimum support $\}$
  end
end
Answer $\cup_k F_k$;

Apriori Algorithm

In the above Fig. 1. Function apriori-gen in line 3 generates Ck+1 from Fk in the following two step process:
1.Join step: Generate RK+1, the initial candidates of frequent

itemsets of size k + 1 by taking the union of thetwo frequent itemsets of size k, Pk and Qk that have the first k□1 elements in common.
Rk+1=Pk U Qk= {item1, item2,.....,itemk, itemk'}
Pk = {item1, item2,.....,itemk, itemk}
Qk = {item1, item2,....., itemk'}
Where item1<item2<…….<itemk<itemk'

2. Prune step: Check if all the itemsets of size k in Rk+1 are frequent and generate Ck+1 by removing those that do not pass this requirement from Rk+1. This is because any subset of size k of Ck+1 that is not frequent cannot be a subset of a frequent itemset of size k +1. '
Function subset in line 5 finds all the candidates of the frequent itemsets included in transaction t. Apriori, then, calculates frequency only for those candidates generated this way by scanning the database. It is evident that Aprioriscans the database at most kmax+1 times when the maximum sizeof frequent itemsets is set at kmax.[3]

.

CONCLUSION

To overview of exploration ,modeling ,deployment and application of different algorithm like nearest neighbor algorithm, Navie algorithm, and Appori algorithm are presented in a simple way the algorithm and method not only used to optimized the problem but also to solve the constrained problem which are dramatically changing just like earthquake. The k nearest algorithm is most reliable and effective algorithm among above discuss algorithms.

REFERENCES

[1] XindongWu • Vipin Kumar & et.al. (2008),'Top 10 algorithms in data mining Knowl Inf Syst',14:1–37DOI 10.1007/s10115-007-0114-2.
[2] Tapas Ranjan Baitharu et.al.,' A Survey on Application of Machine Learning Algorithms on Data Mining', International Journal of Innovative Technology and Exploring Engineering.
(IJITEE) ISSN: 2278-3075, Volume-3, Issue-7, December 2013
[3] Raj Kumar &Dr. Rajesh Verma Classification Algorithms for Data Mining: A Surve y ISSN: 2319 – 1058.
[4] Arun K.Pujari(2013), Data mining Techniques (Third edition), Universities Press (India) Pvt. Ltd., ISBN 978 81 7371 884 7.
[5] Ian H.Witten & et.al. (2013) , DATA MINING Practical Machine Learning Tools and Techniques(Third Edition), MORGAN KAUFMANN PUBLISHERS,ISBN 978-93-80501-86-4. y ISSN: 2319 – 1058.
[6] Rok Rupnik et.alThe Deployment of Data Mining into
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

AUTHOR'S PROFILE

Prof: B. P. Deshmukh
M.Sc. (Comp. science), M.C.M.
Assistant Professor, Department of BCA,
H.V.P.M's D.C.P.E Amravati.

Prof: A. D. Chavan
M.C.A.
Assistant Professor, P.G. Department of Computer Science & Tech. ,
H.V.P.M's D.C.P.E Amravati.