International Research Journal of Nature Science and Technology (IRJNST) E-ISSN: 2581-9038

Volume: 01 Issue: 05 November to December 2019 www.scienceresearchjournals.org

Survey Paper on Data Mining Techniques

N.Ramya[1]
Department of computer science
Bharathiyar University,
Tamilnadu, India

S.Iswarya[2]
Department of computer science
Bharathiyar University,
Tamilnadu, India

Ms E Kanimozhli[3]
Assistant Professor
Department of computer science
Bharathiyar University, Tamilnadu, India

Abstract-

Data mining techniques are used. The statistical techniques is not considered as a data mining technique by many analysts. In this paper, data mining based on clustering, classification, hybridization, regression is researched in detail and the key technology and ways to achieve the data mining areas also surveyed.

Keywords - *Data mining**,** Regression, Hybridization, Classification, Clustering.*

I. INTRODUCTION

Data mining is a set of large amount of data. In the last ten to twenty years, as the volume of stored digital data the memory capability and the completely power have grown, also has the need to take all that potential. For instances. In many industries like communication or retail distribution (e.g: supermarket). There are huge databases of operational data that have plenty of hidden underlying information. The aim of data is uncover that information and provide the decision makes with the knowledge to makers with the knowledge to make better informed decisions. In an academic environment, as is the case of this thesis, the aim is identical, it is to perform knowledge discovery in a huge database.

REGRESSION

Regression technique is a statistical tool for the investigation of relationships between variables. For example, the effect of changes in the money supply upon the inflation rate. To explore such issues, data are assembled on the underlying variables of interest anthem regression techniques are applied estimate the quantitative effect of the casual variables upon
the variable that they influence.

Let us consider an example to identify and quantify the factors that determine earnings in the labour market, with variables like occupation, age, experience, educational qualification, motivation innate ability come to mind, race and gender etc....Let's, education be the single factor for our analysis . Regression techniques with the single explanatory variable is termed as simple regression. In general, regression analysis models the relationship between one or more response variable (also called Dependent variables, explained variables, predicted variables or regress ands)usually named Y and the predictor (also called as independent variables, explanatory variables, control variables or regressors,usually named (X1,X2,.....,XP).Multivariate regression describes models that have more than a one response variance.

MULTIPLE REGRESSION

Multiple regression technique is a statistical procedure that attempts to assess the relationship between a dependent variable and two or more independent variable. When it is assured that more than one independent variables influences the dependent variable substantially, multiple regression technique found to be suitable. The goodness of multiple regression technique can be estimated or predicted by coefficient if determination (R........) which always lies in between 0 and 1. Increasing the number of independent variables, will increase. ... Which appears the regression is improved which is not the case in actual.

Multiple regression analysis uses least squares method to simultaneously estimate the individual effects, or coefficient of each independent variable on the dependent variable. Multiple regression analysis, uses the following general equation, to predict the value of Y (dependent variable) for given values of X1 (1st independent variable), X2 (second independent variable), Xu (Xjth independent variable) as:

$Y = a_0 + b_1x_1+b_2x_2+\ldots+b_jx_j$ + Random term (β)

Where a_0 is the intercept or constant and $b_1,b_2\ldots b_j$ are slope in multiple linear regression, also known as regression coefficient. These coefficient as group tends to minimize the sum of squares of differences between actual and calculated value of Y. Statistical significance of multiple regression coefficients can be tested by F-test through the analysis of variance or t-test which is bases as t-value and its degree of freedom. These test determines whether the estimated coefficient b_1 is sufficiently statistically significant as opposed to one that might just appear to be significant.

It is to be noted that only independent variables that are in use and are actively charging in relationship to varying dependent variables should be considered for possible inclusion in regression technique. For example, the yield of rice per acre depends upon quality of seeds, fertility of soil, fertilized used, rainfall, temperature etc.

Further, testing using various additional independent variable is also a possibility with regression of "Y on a single independent variable is important but the details of y is found to have unexplained. For example, adding labour hours an then direct labour rupees/costs, which depends upon labour hours, is not appropriate as these independent variables are essentially the same.

In case of multiple linear regression, a condition referred to as multicollinearity to have some kind of correlation among two or more of the independent variables, hardly and correlation between them exists, In that situation of minimal multicollinearity, the regression's line of best fit to the dependent variable ils not significantly affected but causes regression coefficient to vary radically, unexpectedly, and produce similarly valid but ineffective results.

KERNEL BASED REGRESSION

Kernel based regression technique involves weighting each neighbouring data point according to a kernel function giving a decreasing weight with a distance and then computing a weighted local mean or linear or polynomial regression model. In this way, this method looks similar to smoothing interpolation as in geographic space, of the predictor variables. The primary tuning parameter is the bandwidth of kernel function, each is relatively specified in such a way that the same value can be applied along all predictor axes. Larger the bandwidth, the smoother will be the functions. Hence, the form of the kernel function becomes secondary importance.

PARAMETRIC AND NON PARAMETRIC REGERSSION

Parametric regression is one in which regression function is defined in terms of a finite number of unknown parameters that are estimated from the data, where as non-parametric regression refers to a technique where regression coefficient lies in a specified set of function, which might be infinite -dimensional. In parametric regression technique, assumption of data are to be linear.

However, in non-parametric case, mean of observed values are taken into consideration, for each case. Non parametric regression typically assumes kittle else about the shape of the regression function beyond some degree of smoothness, which make this technique more durable inferences that will be an indispensable tool for many researchers. In order words, the data analysis with non-parametric method provides a greater ability to inference procedures that are less dependent on functional assumptions. Hence, if one has some reservations in a parametric form, then alternate non parametric specification tests may provide reassurance to the data analyst.

Non-parametric method with all its strengths, it has some concerns like: complex, computationally intensive, hence in need of large data sets, as relationships are observed by examining nearby observations, and finally unavailability of software to handle unified non-parametric models automatically.

HYBRIDIZATION

Hybrid systems employ more than one technology to solve a problem with more efficiency. This can be used when its application provides a better solution, in comparison to the individual ones. Further, the hybridization of technologies can have some pitfalls, hence need utmost care before successful implementation.

Classification of hybrid system

The following are some of the ways to classify the hybrid systems

- Embedded hybrid system, where the data mining techniques are appeared to be fused totally.
- Sequential hybrid system in which the data mining techniques are to be used in a pipeline fashion.
- Finally, auxiliary hybrid system where one technique calls the other as subroutine.

We may consider the application of soft computing methods (Neural network (NN), fuzzy logic (FL) and genetic algorithm (GA)) to develop an efficient hybrid system.

In embedded hybrid system, we may use neural network and fuzzy logic to fuse completely, in which neural net supposed to receive the fuzzy inputs, then process it and finally extracts the fuzzy outputs

NN-FL Hybrids: The integration of NN and FL systems found to be useful in

- Accompanying mathematical relationships among many attributes in a dynamic way.
- Performance mapping is obtained is obtained with some degree of imprecision.

The NN-FL hybridization can be obtained to provide an efficient data mining application with the following ways:

- Neural learning capabilities can be applied to fuzzy learning systems, such as to make FL systems to be more adaptive to the dynamic scenarios. This is popularly called as NN driven fuzzy reasoning;
- The other way is to provide NNs with fuzzy capabilities, so that the flexibility and expressiveness of the network will be enhanced suitable for uncertain environments.

NN- GA Hybrids: The hybridization of Gas and NNs have the ability to locate neighbourhood of the optimal solutions faster than any other available conventional search methods. The limitation of this hybrids include requirement of large memory for handling chromosomes for a given network under investigation and the network scalability issues, due to large networks.

FL-GA Hybrids: In this case, the several fuzzy logic parameters such as input variables and membership function have been optimized using GAs.

CLASSIFICATION

Classification, if the accuracy of the model is considered to be acceptable. Such data are refer to in the machine learning or data mining literature as “precisely unseen” For classification rules learned from the analysis of data from existing customers can be used to predict the credit rating of new or future customers.

Model construction:

The model is constructed in a step-by-step manner as follows:

- Each data samples or tuples or examples or objects are assumed to belong to a predefined class, as determined by one of the attributes, called as class label,
- All the data samples used for the model construction is called as Training data set,
- Since, the class label of each training sample of provided, the learning is called as supervised learning. In contrast, in unsupervised learning, the class label of each training samples are known and the number or set of classes to be learned may not be known well in advance,

Finally, the model is constructed and represented in the following forms:

- Classification rules, IF-THEN statements;

- Decision tress;
- Mathematical formula, etc.

Built model used for classifying feature or unknown objects:

In this, the following points are to be remembered

- Accuracy rate is defined as the percentage of test set samples that are correctly classified by the model, and
- Test set is preferable different than training set so as to avoid over-fitting which may have incorporated some particular anomalies of the training data that are not present in the overall sample population, will occur.
- Finally, if the accuracy of the model is acceptable, then the built model can be used to classify the future objects for which the class label is not known.

Classification methods:

Decision tree induction has been studied in both the areas of pattern recognition and machine learning, which synthesizes the experience gained by the researchers in the area of machine learning to describe a computer program called ID3.

ID3 Algorithm:

This is considered to be one of the famous inductive logic programming methods developed by Quinlan. It is essentially an attribute based machine learning algorithm that constructs a decision tree based on training set of data and a entropy measure to build the leaves of the tree. The informal formulation of ID3 is as follows:

Step1: Determine the attribute that has the highest information gain on training set.

Step2: Using the said attribute as the root of the tree, create a branch for each of the values that the attribute can take.

Step3: Finally, for each of the branches, repeat this process with the subset of training set is classified by tis branch.

Draw backs of ID3:

- ❖ It requires a lot computation at every stage of construction of decision tree;
- ❖ It also needs all the training data to be in the memory; and
- ❖ It does not suggest any standard splitting index for range attributes

C4.5 Algorithm:

C4.5 is an algorithm used to generate a decision tree developed by Rose Quinlan and popularly considered as an extension to earlier ID3 algorithm. This is sometimes referred to as a statistical classifier. The following are some of the improvements that C4.5 are in comparison to ID3.

- o Handling training that with missing attribute values, where C4.5 allows attribute values to be marked for missing which are not used for calculation of gain and entropy.
- o C4.5 handles attributes with differing costs.

J48 Algorithm:

J48 is based on the ID3 algorithm developed by Ross Quinlan, with additional features to address problems that ID3 was unable to deal. In practice, C45 uses one successful method for finding high accuracy hypotheses, based on pruning the rules issued from the tree constructed during the leaning phases.

However, the principal disadvantage of C4.5 rule sets is the amount of CPU time and memory they require.

Given a set of cases, J48 first grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the cases in S belong to the same class or S is small, the tree is leaf labelled with the most frequent class in S.
    - o There are usually many test that could be chosen in this last step

- After the building process, each attribute test along the path from the root to the leaf becomes a rule antecedent and the classification at the leaf node becomes the rule consequence.

To illustrate the post of the rules, let us consider the following rule generated from the tree:

IF (service=login) ^ (flag=SF)
THEN class= ftp_ write

Instead of E/N, J48 algorithms determine the upper limit of the binomial probability when E events have been observes in N trials, using a user-specified confidence whose default value is 0.25.

Pros and cons of decision trees:

Decision trees in general have several important advantages:

- Extremely fast at classifying unknown records.
- Easy to interpret for small-sized trees.
- Work well in the presence of redundant attributes.
- If methods for avoiding over fitting are provided then decision trees are quite robust in the presence of noise.
- Robust to the clear indication of which fields are most important for prediction.
- It can handle both nominal and numeric input attributes.
- Capable of handling data-sets that may have errors.
- Capable of handling data-sets that may have missing values.
- Non-parametric method, it means decision trees have no assumptions about the space distribution and the classifier structure.

disadvantage as:

1. Most of the algorithms require that the targets attribute will have only discrete values.
2. The greedy characteristic of decision tress leads to another disadvantage that should be pointed out.

CLUSTERING:

Clustering is considered to be the most unsupervised learning problem in which it deals with finding a structure in a collection of unlabelled data. Further, it can be loosely defined as the process organizing objects into graphs whose members are similar in some way. *N this perspective, a clus*ter can be understood as a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to the other clusters. Traditional clustering techniques can be categorized into two groups as **partition and hierarchical.**

PARTITION CLUSTERING:

Hard Partitioning

This kind of methods divides a data set strictly into disjoint subsets. Conventional clustering algorithm find a “hard partition” of a given data set based on certain criteria that evaluate the goodness of partition. More formally, we can define the concept of “hard partition” as follows:

(1) Let X be a data set of data, and x_i be an element of X. A partition p= {C_1, C_2,…C_j} of X is “hard” is and only if

(i). $\forall\ x_1 \in x\ \exists\ C_j \in P$ such that $x_i \in C_j$

(ii). $\forall\ x_i \in X,\ x. \in C_j \rightarrow x_i \in C_k$

This condition in the definition assures that the portion covers all data points in X; the second condition assures that all the clusters in portion are manually exclusive.

HIERARCHICAL CLUSTERING ALGORITHMS

Working principle: Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic processes of hierarchical clustering (defined by S.C. Johnson in 1967) are:

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distance between the clusters the same as the distance between items by contain.

2. Find the closest pair of clusters and merge then into a single cluster, so that now you have one cluster less.
3. Compute distance between the new cluster and each of the old clusters.
4. Repeat step 2 and 3 and till all items are clustered in to a single cluster of size N.

Step 3 can be done in different ways, which distinguishes single-linkage from complete-linkage and average-linkage clustering.

A variation on average-link clustering is the UCLUS method of R. D Andrade (1978) which uses the median distance, which is much more outlier-proof than the average distance .There is also a divisive hierarchical lustering which does the reverse by starting with all objects in. one clusters and subdividing them into smaller pieces. Divisible methods are not generally available, and really have been applied.

Further, it is worth noting that when there is no point in having all the N items grouped in a single cluster but, once you have got the complete hierarchical tree, if we want k clusters you just have to cut the k-longest links.

Single-Linkage Clustering:

The Algorithm

Let's now take a deeper look at how Johnson's algorithm works in the case of single-linkage clustering.

The N*N proximity matrix is D = [d(i,j)]. The clustering are assigned sequence numbers 0,1,…(n-1) and L(k) is the level of the kth clustering

The algorithm is composed of the following steps:

1. Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

Increment the sequence number: m = m+1rUpdate the proximity matrix, D. by deleting the rows and columns corresponding to clusters (r) and (s) and a row and a column corresponding to the newly formed cluster.

Problems:

The main weaknesses of agglomerative clustering methods are: They do not scale well time complexity of at least O (n^2), where n is the number of total objects;They can never undo what was done previously.

II. CONCLUSION

There are several techniques in data mining each and every technique has its advantage and disadvantageThe commercial, educational and scientific applications are increasing by dependent on these methodologies.

REFERENCE

[1] Alek Kumar Awasthi Department of Computer Science and Engineering, Sarasvati Higher Education and Technical College of Engineering, Varanasi, Uttar Pradesh, India; Study of Data Mining Techniques and Its Types, ISSN: 2456 – 3307

[2] Nikita Jain1, Vishal Srivastava2 1M. Tech. Scholar, 2Associate Professor, Arya College of Engineering and IT, Rajasthan, India, nikitagoodjain@gmail.com, vishal500371@yahoo.co.in; DATA MINING TECHNIQUES: A SURVEY PAPER,eISSN:2319-1163|pISSN:2321-7308

[3] Smita1 , Priti Sharma2 1 (Student, M. Tech, Amity University) 2 (Assistant Professor, Amity University); Use of Data Mining in Various Field: A Survey Paper, EISSN :2278-0661|pISSN : 2278-8227

[4] Anoop Kumar Jain1 and Satyam Maheswari Dept. of Computer Application Samrad Ashok Technological Institute, Vidisha (M.P.), India anoopjain0108@gmail.com Dept. of Computer Application Semrad Ashok Technological Institute, Vidisha (M.P.), India satyam.vds@gmail.com; Survey of Recent Clustering Techniques in Data Mining ISSN: 0976-4828

[5] K.Kameshwaran1 , K.Malarvizhi2 1 M.E-CSE, Department Of Computer Science & Engineering, Coimbatore Institute of Technology Coimbatore, Tamil Nadu, India. 2 Associate Professor in CSE, Department Of Computer Science & Engineering, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.; Survey on Clustering Techniques in Data Mining ISSN : 0975-9646