IJCSC International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 373-378

A Perception of Statistical Inference in Data Mining

Sanjay Gaur[1] & M. S. Dulawat[2]

[1-2]Department of Mathematics & Statistics, Mohanlal Sukhadia University, Udaipur, India
Email: [1]sanjay.since@gmail.com, [2]dulawat_ms@rediffmail.com

ABSTRACT

As we know that data mining is concern with learning from data therefore, completeness, quality and real world data preparation, is a key prerequisite of successful data mining with its aim to discover something new from the facts already recorded in the certain database. Preparation of data is a fundamental stage of data analysis. During data preparation, the major problem occurs due to missing values, impure values and outliers. To overcome this situation, some of the statistical techniques are required to apply during the data preparation. Therefore, erroneous data may be corrected and removed, whereas missing data must be supplied or estimated. This is one of the important parts of the data mining, which comes in the scene with the help of the statistical inference. The aim of the statistical techniques is to understand the patterns of correlation and causal links among the data values which is explained or making predictions for future data values as generalization.

Keywords: Data Mining, Statistical Inference, Data Preprocessing, Data Transformation, E M Algorithm, Missing Values

1. INTRODUCTION

The area of data mining is concern with learning from data and turning or transforming data into meaningful data means information. This information is base of decision for any level of management and organization. Data always remain in the database, thus all the final reports are generated by the help of the database. If our database is impure then our final reports are automatically going to draw wrong picture. Here the problematical area in the database is impure data and missing values.

Data in the database with missing values and impure values complicates both analysis and application of a solution to the set of new data or data mining. Preparation of data is a fundamental stage of data analysis. During data preparation, the major problem occurs due to missing values, impure values and outliers. To overcome this situation some of the statistical techniques are required to apply during the data preparation. With the help of statistical methods and techniques, we can recover the incompleteness of missing values and reduce ambiguities. Therefore, statistical inference is now subject to prepare model about how to transform the data into complete and cleaned form.

The role of statistical methods and inference has gained importance in exploring estimation and prediction techniques. Wilks [1]have considered estimation of parameters of a normal bivariate population with missing values.

Latter on, Edgett [2], Anderson [3] and Nicholson [4] found the maximum likelihood solution for the parameters of normal population. Buck [5] suggested estimation of missing values in multivariate data suitable for use with an electronic computer. Dempster, Laird and Rubin [6] studied about maximum likelihood estimation from incomplete data via EM algorithm. Kim and Curry [7] considered the treatment of missing data in multivariate analysis. Rubin [8], [9] explored about inference and missing data and multiple imputations for non-response in the survey. Ramchandran [10] gave a detailed study on univariate, bivariate and multivariate analysis with missing values. Clark, Madigan, Pregobon and Smyth [11], [12] studied and explained the relation between statistical method and data mining. David [13], [14] considered that the disciplines of statistics and data mining have common aims.

Allison [15], [16] investigated estimates of linear models with incomplete data and on missing data. Smyth [17], Zhang, Zhang and Young [18] have considered that data preparation is a fundamental stage of data analysis. Chen, Drane, Valois and Drane [19] studied and discussed about multiple imputation for missing ordinal data. Qin [20] considered the semi-parametric optimization for missing data imputation. Recently Nisbet, Elder and Miner[21] published book on statistical analysis and data mining which deals with massive and complex datasets with novel statistical approaches to evaluate, analyze and to draw the conclusions.

2. KEY PROCESS OF DATA MINING

Data preprocessing, cleaning and transformation is an important and key part of data mining. The primary data

or some time secondary data are always found dirty and not ready for data mining in the real world. They are suffering with non-integration, error, inconsistency, noise and missing values etc. Therefore, statistical methods and inference plays important role to make database pure and meaningful. Data mining includes classification, estimation, prediction, clustering, association rules and visualization. Classification, estimation and prediction are directly associated with statistical mathematics and inferential statistics. Thus task of statistics are well suited for data mining. Data recorded in the database for data mining process are obtained from different and heterogeneous sources. Thus, the data used by the process may have impure, incorrect or missing data. Therefore, erroneous data may be corrected and removed, whereas missing data must supplied or estimated.

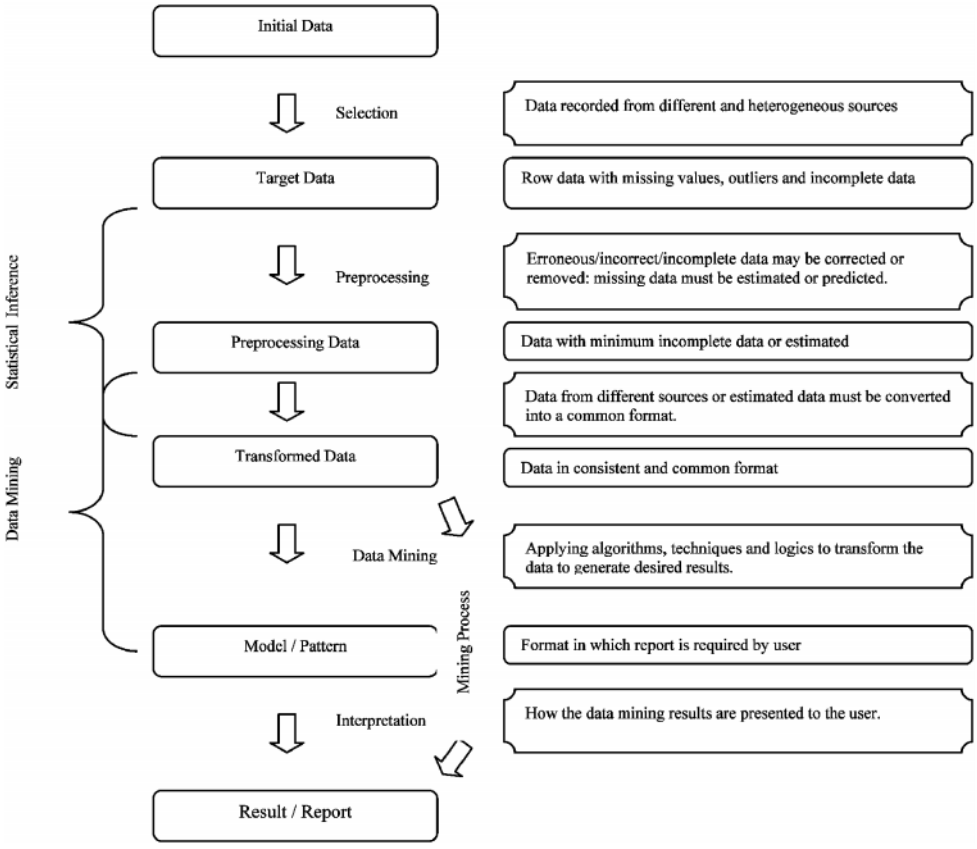

Fig.: Steps of Data Mining Process

This is one of the important parts of the data mining, which comes in the scene with the help of statistical inference. This stage of mining is known as the preprocessing of data. The next action is to convert estimated data into a common format for processing because data are collected from different sources. Now the base for data mining is ready, here we can apply algorithm to transform the data as our requirement. Transformation of data make database easier to mine and efficient to provide smooth processing of algorithm.

3. CHALLENGES TO MANAGE MASSIVE DATASETS

Technical advances led to new and automated data collection and refinement methods. Datasets, once at a premium are often, overflowing nowadays and sometimes indeed massive. A new breed of challenges is the need for methodology to analyze such masses of data with a view to understanding complex phenomena and relationships. Such capability provided by data mining which combines core statistical techniques with machine intelligence.

An oft-stated goal of data mining is the discovery of patterns and relationships among different variables in the database. This is not different from some of the goals of statistical inference: consider for instance, simple linear regression.

Both statistics and data mining are concerned with drawing inferences from data. Data mining attracts such problems as obtaining efficient summaries of large amounts of data, identifying interesting structures and relationships within a data set, and using a set of previously observed data to construct predictors of future observations. Statisticians have well established techniques for solving all of these problems. Many statistical models exist for explaining relationships in a data set or for making predictions.

Massive datasets can be tackled by the sampling (if the aim is molding, but not necessarily if the aim is pattern detection), by adoptive methods, or by summarizing the records in terms of sufficient statistics.

Various problems arise from the difficulties of accessing very large data sets. The statistician's conventional view point of a flat data file in which rows represents object and columns represents variables, may bear no resemblance to the way the data are stored. In many cases the data are distributed, and stored on many machine. Obtaining a random sample from the data that split up in this way is not a trivial matter. Data mining is typically a secondary process of data analysis that is, the data originally collected for some other purpose.

4. APPROACHES OF STATISTICAL METHODS IN DATA MINING

Statistical inference concept like determining a data distribution and calculating a mean and variance viewed as data mining techniques. These techniques in an integration form are known as model for mining. In general, data mining modeling process requires searching the fact by the help of the statistical methods and statistical inference. An important part requires inference for the result of the search applicable to general model. There may be different methods for data mining, these method are design about the specific type of problems and require specific type of data structure within an algorithm approaches, means a clear stated modeling to solve the problem.

The statistical approaches are associated with the relationship between input and output which is data driven. Thus, purity of database plays important role when all the results are data driven. The statistical inference based modeling is suitable for database application with large amount of dynamically changing data.

4.1. Samples and Probability in Dataset

Many data mining problem involves the entire population of interest while other involves just a sample from a population. In other case, even though the complete dataset is available, the data mining operation carried out on a sample.

Statistical inference allows us to make statement about population structure, to estimate the size of these structures and to state our degrees of confidence in them, all based on the sample.

In order to make an inference about a population, we must have a model or pattern structure in the mind: we would not able to assess the evidence for some structure underlying the data if we never contemplated the existence of such a structure.

Statistical inference is based on the promises that the sample has been drawn from the population in the random manner- that each member of the population had a particular probability of appearing in the sample. The model will satisfy the distribution function for the population – the probability that a particular value for the random variable will arise in the sample.

Let $P(x\ (i))$ be the probability of individual i having vector measurement $x(i)$ (here p could be a probability mass function or a density function, depending on the nature of x). If we further assume that the probability of each member of the population being selected for inclusion in the sample has no effect on the M probability of other members being selected (that is, the separate observation are independent, or that the data are drawn" at random"), the overall probability of observing the entire distribution of values in the sample is simply the product of the individual probabilities:

$$P(D|\theta, M) = \prod_{i=1}^{n} P(x(i)|\theta, M)$$

Where M is, the model and θ are the parameters of the model, which assumes as fixed at this point). When regarded as a function of the parameter θ in the model M. this is called the likelihood function.

4.2. Point Estimation

Estimation of population parameter gives by a single number called a point estimation of the parameter. An estimate of parameter gives by two numbers between which the parameters may be considered to lie is called and interval estimate of the parameter. Interval estimate indicates the precision or accuracy of an estimate and are therefore preferable to point estimate.

Point estimation refers to the process of estimating a population parameter, θ, by an estimate of the parameter, θ. This can be done to estimate mean, variance, standard deviation, or any other statistical parameter. Often the estimate of the parameter for a general population may made by actually calculating the parameter value for a population sample. An estimator technique may also used to estimate (predict) the value of missing data. The bias of an estimator is the difference between the expected value of the estimator and the actual value.

$$\text{Bias} = E(\hat{\theta}) - \theta$$

An unbiased estimator is one whose bias is 0.while point estimator for small data sets may actually be unbiased, for larger database applications we would expect that most estimators are biased.

4.3. Mean Square Error

One measure of the effectiveness of an estimate is the mean square error (MSE), which defined as the expected

value of the squared different between the estimate and the actual value:

$$MSE(\hat{\theta}) = E((\hat{\theta}-\theta)^2$$

The squared error often examined for a specific prediction to measure accuracy rather than to look at the average difference. The squaring performed to ensure that the measure is always positive and to give a higher weighting to the estimates that are grossly inaccurate. The MSE commonly used in evaluation data mining prediction techniques. At times instead of predicting a simple point estimate for a parameter, one may determine a range of value within which the true parameter value should fall. This range is called confidence interval.

4.4. Root Mean Square

The root mean square (RMS) may also be used to estimate error or as another statistics to describe a distribution. Calculating the mean does not indicating the magnitude of the value. The RMS can be used for this purpose. Given a set of n value $X = \{x_1, x_2, \ldots, x_n\}$, the RMS is defined by

$$\text{RMS} = \sqrt{\sum_{j=1}^{n} x_j^2 \Big/ n}$$

An alternative use is to estimate the magnitude of the error. The root mean square error (RMSE) is found by taking the square root of the MSE. A popular estimating technique is the jackknife estimate. With this approach, the estimate of parameter, $\hat{\theta}$, is obtained by omitting one value from the set of observed values. Suppose that there is a set of n value $X = \{x_1, x_2, \ldots, x_n\}$. An estimate for the mean would be

$$\hat{\mu}_{(i)} = \left(\sum_{j=1}^{i=1} + \sum_{j=i+1}^{n} x_j \right) \Big/ (n-1)$$

Here the subscript (i) indicates that this estimate is obtained by omitting the i^{th} value. Given a set of jackknife estimates $\hat{\theta}_{(i)}$, these can in turn be used to obtain an over all estimate

$$\hat{\theta}_{(.)} = \sum_{j=1}^{n} \hat{\theta}_{(j)} \Big/ n$$

4.5. Maximum Likelihood Estimate

Another technique for point estimation is called the maximum likelihood estimate (MLE). Likelihood can be defined as a value proportional to the actual probability that with a specific distribution the given sample exists. Therefore, the sample gives us an estimate for a parameter from the distribution. The higher the likelihood value, the more likely the underlying distribution will produce the result observed.

A sample set of values $X = \{x_1, x_2, \ldots, x_n\}$ from a known distribution function $f(x_i|\theta)$, the MLE can estimate parameter for the population from which the sample is drawn. The approach obtain parameter estimate that maximum the probability that the sample data occur for the specify model. It looks at the joint probability for observing the sample data by multiplying the individual probability. The likelihood function L is thus defined as

$$L(\Theta|x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} f(x_i|\Theta)$$

The value Θ of that maximizes L is the estimate chosen. This can be found by tacking the derivate (perhaps after finding the log of each side to simplify the formula) with respect to Θ.

4.6. Expectation Maximization (EM) Algorithm

The expectation-maximization (EM) algorithm is an approach that solves the estimate problem with incomplete data. The EM algorithm find an MLE for a parameter (such as mean) using a two step process: estimation and maximization. The basic EM algorithm is as

Input:

$\Theta = \{\theta_1, \ldots, \theta_p\}$ // Parameters to be estimated

$X_{obs} = \{x_1, \ldots, x_k\}$ // Input database values observed

$X_{miss} = \{x_{k+1}, \ldots, x_n\}$ // Input database values missing

Output:

$\hat{\Theta}$ // Estimates for Θ

EM algorithm:

i: = 0;

Obtain initial parameter MLE estimate, $\hat{\Theta}^i$;

repeat

Estimate missing data, $\hat{X}^i_{miss}$;

i++

Obtain next parameter estimate, $\hat{\theta}^i$ to

maximize likelihood;

until estimate converges;

An initial set of estimates for the parameter is obtained. Given this estimation and the training data as input, the algorithm then calculates a value for the missing data. For example, might use the estimated mean to predict a missing value. These data (with the new value added) are then used to determine an estimation for the mean that maximize the likelihood. These steps

are applied iteratively until successive parameter estimate converge. Any approach can be used to find the initial parameter estimates. In above mentioned algorithm it is assumed that the input database has actual observed values $X_{obs} = \{x_1, ..., x_k\}$ as well as value that are missing $X_{miss} = \{x_{k+1}, ..., x_n\}$. We assume that the entire database is actually $X = X_{obs} \cup X_{miss}$. The parameters to be estimated are $\Theta = \{\theta_1, ..., \theta_p\}$. The likelihood function is defined by

$$L(\Theta \mid X) = \prod_{i=1}^{n} f(x_i \mid \Theta)$$

We are looking for the Θ that maximize L. the MLE of Θ are the estimate that satisfy

$$\frac{\partial ln L(\Theta \mid X)}{\partial \theta_i} = 0$$

The expectation part of the algorithm estimates the missing value using the current estimates of Θ. This can initial be done by finding a weighted average of the observed data. This maximization step then finds the new estimates for theparameters that maximize the likelihood by using those estimates of the missing data.

4.7. Bayes Theorem

With statistical inference, information about a data distribution are inferred by examining data that follow which distribution. For set of data $X = \{x_1, ..., x_n\}$, a data mining problem is to uncover properties of the distribution from which the set comes. Bayes rule is a technique to estimate the likelihood of a property given the set of data as evidence or input. Suppose that either hypothesis h_1 or hypothesis h_2 must occur, but not both. Also suppose that x_i is an observable event.

$$P(h_1 \mid x_i) = \frac{P(x_i \mid h_1)P(h_1)}{P(x_i \mid h_1)P(h_1) + P(x_i \mid h_2)P(h_2)}$$

Here $P(h_1 \mid x_i)$ called the posterior probability, while $P(h_1)$ is the prior probability associated with the hypothesis h_1. $P(x_i)$ is the probability of the occurrence of data value x_i and $P(x_i \mid h_1)$ is the conditional probability that, given a hypothesis, the tuple satisfies it. Where there are m different hypotheses, we have:

$$P(x_i) = \sum_{j=1}^{m} P(x_i \mid h_j)P(h_j)$$

Thus, we have $P(h_1 \mid x_i) = \frac{P(x_i \mid h_1)P(h_1)}{P(x_i)}$

Bayes rules allows us to assign probabilities of hypotheses given a data value, $P(h_j \mid x_i)$. Here discuss tuples when in actuality each x_i may be an attribute value or other data lable. Each h_1 may be an attribute value (such as range), or even a combination of attribute values.

4.8. Hypothesis Testing

Hypothesis testing attempts to find a model that explain the observed data by first creating a hypothesis and then testing the hypothesis against the data. This is in contrast to most data mining approach, which creates the model from the actual data without guessing what it is first. The actual data it self drive the model creation. The hypothesis usually is verified by examining a data sample. If the hypothesis holds for the sample, it is assumed to hold for the population in general. Given a population, the initial (assumed) hypothesis to be tested, H_o, is called the *null hypothesis*. Rejection of the null hypothesis cause another hypothesis, H_1, called the *alternative hypothesis*, to be made.

One technique to perform hypothesis testing based on the use of the chi-squared statistic. Actually, there is a set of a procedure referred to as chi squared. These procedures can be used to test the association between two observed variable values and to determine if a set of observed variable value is statistically significant. A hypothesis is first made, and then the observed values are compared based on this hypothesis. Assuming that O_i represents the observed data and E_i is the expected value based on the hypothesis, the *chi-squared statistic,* χ^2, is defined as:

$$\chi^2 = \sum_{i=1}^{k} \left((O_i - E_i)^2 / E_i \right)$$

When comparing a set of observed variable to determine statistical significance, the values are compared to those of the expected case. This may be the uniform distribution.

4.9. Regression and Correlation

Both bivariate regression and correlation can be used to evaluate the strength of a relationship between two variables. Regression is generally used to predict future values based on past value by fitting a set of points to a curve. Correlation, however, is used to examine the degree to which the values for two variables behave similarly.

Linear regression assumes that a linear relationship exists between the input data and output data. The common formula for a linear relationship is used in this model.

$$Y = c_0 + c_1 x_1 + ... + c_n x_n$$

Here there are n input variable, which are called predictors or regressors; one output variable (the variable being predicted), which is called the response; and $n + 1$ constants, which are chosen during the modeling process to match the input example (or sample). This is something called multiple linear regression because of more than one predictor.

5. CONCLUSION

Technical advancement comprise led to new and automated data collection and refinement methods. The goal of data mining is to discover the patterns and relationship among different variables in the database. This is just similar to the goal of the statistical inference like simple line regression, correlation etc. Both statistical inference and data mining are concerns with the drawing inference from data with the objective of understand patterns, correlation and casual links between variables. It is found that statistical methods are very much suitable for estimation and prediction for database values. It may be applicable on primary data collection as well as secondary data refinement and prediction. Therefore statistical methods or inference in context of data mining closely and completely concern with the estimation and prediction. The basic statistical methods like Point Estimation, Maximum Likelihood Estimation, EM Algorithm, Hypothesis Testing, Regression and Correlation etc are always positively contribute toward the data mining.

REFERENCES

[1] S.S.Wilks, "Moment and Distribution of Estimates of Population Parameters from Fragmentary Samples", *Annals of Mathematical Statistics*, **3**, pp.163-165, 1932.

[2] G.L. Edgett, "Multiple Regressions with Missing Observation Among the Independent Variables", *J. American Statistical Association*, **51**, pp.122-131, 1956.

[3] T.W. Anderson, "Maximum Likelihood Estimation for a Multivariate Normal Distribution when Some Observation are Missing", *J. American Statistical Association*, **52**, pp. 200-204, 1957.

[4] A. Nicholson, "Estimation of Parameters from Incomplete Multivariate Sample", *J. American Statistical Association*, **52**, pp.523-526, 1957.

[5] S.F. Buck, "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer", *J. Royal Statistical Society*, Series B, **2**, pp.302-306, 1960.

[6] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likelihood Estimation from Incomplete Data Via EM Algorithm", *J. The Royal Statistical Society*, Series- B, **39**, pp.1-38, 1977.

[7] J.O. Kim, and J. Curry, "The Treatment of Missing Data in Multivariate Analysis", *Social Methods and Research*, **6**, pp. 215-240, 1977.

[8] D.B. Rubin, "Inference and Missing Data", *Biometrika*, **63**, pp. 581-592, 1976.

[9] D.B. Rubin, *Multiple Imputation for Non-response in Surveys*, John Wiley and Sons, New York, 1987

[10] P. Ramchandran, "Missing Values: Alternatives in Data Analysis", *Department of Research Methodology, Tata Institute of Social Sciences*, Deonar, Bombay, Series No. 61, 1987.

[11] G. Clark, D. Madigan, D. Pregibon, and P. Smyth, "Statistical Inference and Data Mining, Communication of ACM", **39**, pp. 35-41, 1996.

[12] G. Clark, D. Madigan, D. Pregibon and P. Smyth, "Statistical Themes and Lessons for Data Mining", *Data Mining and Knowledge Discovery,* Kluwar Academic Publisher, Manifacutered in Netherlands, **1**, pp. 11-28, 1997.

[13] J.H. David, "Data Mining: Statistics and More? The American Statistician, **52(2)**, pp. 112-118, 1998.

[14] J.H. David, "Statistics and Data Mining : Interesting Disciplines", *Department of Mathematics, Imperial College*, London, UK, SIGKDD Exploration, **1(1)**, pp. 16-19, 1999.

[15] P.D. Allison, "Estimation of Linear Models with Incomplete Data", *Social Methodology*, San Francisco: Jossey Bass, pp. 71-103, 1987.

[16] P.D. Allision, *Missing Data*, Thousand Oaks CA: Sage Publication, 2001.

[17] P. Smyth, "Data Mining at the Interface of Computer Science and Statistics", *Data Mining for Scientific and Engineering Applications,* Department of Information and Computer Science, University of California, CA, 92697-3425, Chapter 1, pp. 1-20, 2001.

[18] S. Zhang, C. Zhang and Q. Young, "Data Preparation for Data Mining", *Applied Artificial Intelligence*, **17**, pp. 375-381, 2003.

[19] L. Chen, M.T. Drane, R.F. Valois and J.W. Drane, "Multiple Imputation for Missing Ordinal Data", *Journal of Modern Applied Statistical Methods*, **4(1)**, pp. 288-299, 2005.

[20] Y.S. Qin, "Semi-parametric Optimization for Missing Data Imputation", *Applied Intelligence*, **27(1)**, pp.79-88, 2007.

[21] R. Nisbet, J. Elder and G. Miner, *Statistical Analysis & Data Mining Applications*, Academic Press, An Imprint of Elsevier, New York, 2009.