ORIGINAL RESEARCH PAPER

Computer Science

A DETAIL INVESTIGATION ON THE MEDICAL DATABASES BY IMPLEMENTING VARIOUS METHODS

KEY WORDS:

Hiren R. Kavathiya Assistant Professor Department Of Computer Science & I.T. Shree M.V. & N.V. Virani Science College, Rajkot

Dr. G. C. Bhimani* Professor & Head, Department Of Statistics Saurashtra University, Rajkot *Corresponding Author

ABSTRACT

Outlier detection is presented in detail in chapter 1.The finding of outliers for high dimensional datasets is a challenging data mining task. Different perspectives can be used to define the notion of outliers. Hawkins et al., 2002, defines an outlier as "an observation which deviates so much from other observations as to create suspicions that it was generated by a different mechanism". While 'Barnett and Lewis, 1994' define it as "An outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that dataset".

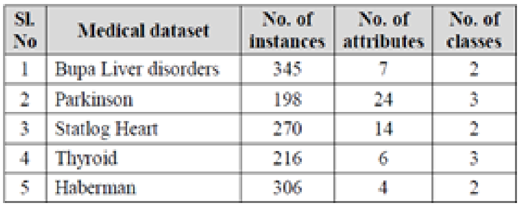
STATISTICAL OUTLIER DETECTION METHODS

Previously, the outlier detection issues were resolved by the creation of a data model based on probability and also making use of mathematical methods of applied statistics and probability theory. A probabilistic model can be either a priori given or automatically constructed by the given data. Any data object whether belonging to probabilistic model or any other distribution law is decided from the construction of the probabilistic model and in any case that it doesn't suit that model, it is measured as an outlier. Standard probability distributions and combinations of the same are used to create probabilistic models and at times, they may include unknown parameters projected while data mining. Algorithms for approximating probability distributions by empirical data exist along with a priori given probability distributions.

The condition when it comes to the regression model is much more complicated. Here, certain outlying points will influence the regression much more than certain others. In massive volumes of highly multidimensional data (which makes the task challenging), it is of great importance to detect outliers while data mining. The challenging nature is due to the fact that crucial outliers may hide in one dimensional data vision. This makes the detection of one dimensional outliers based on scanning one field or variable or attribute at a time ineffective. There are three fundamental approaches to the problem of outlier detection:

1. Type 1 – The outliers are estimated without any previous knowledge of the data. This is essentially a learning approach analogous to unsupervised clustering. The approach processes the data as a static distribution, pinpoints the most remote points, and flags them as potential outliers.
2. Type 2 – Represents normality as well as abnormality. This approach requires pre-labelled data, tagged as normal or abnormal.
3. Type 3 - Models only normality (or in a few cases models, only abnormality). It may be considered semi-supervised as although the normal class is taught, the algorithm learns to recognize abnormality as well.

Outlier detection methods can be divided into **uni variate** and **multi variate** methods.



Sl. No	Medical dataset	No. of instances	No. of attributes	No. of classes
1	Bupa Liver disorders	345	7	2
2	Parkinson	198	24	3
3	Statlog Heart	270	14	2
4	Thyroid	216	6	3
5	Haberman	306	4	2

Above table represents a general design of an outlier detection technique. The healthcare industry has become the new hub for applying the evidence based medicine approach based on the combination of data warehousing and data mining. Detailed data on evidence based medical applications are in demand by all parties. Data has more importance today than any other approach. And many times, although the data might be available but valuable information might not be.

MODEL 1: APPLICATION OF DATA MINING TECHNIQUES FOR OUTLIER MINING IN MEDICAL DATABASES

INTRODUCTION

The subject of outlier detection has been very crucial for data analysis and any complicated issues in terms of patient symptoms, diagnoses and behaviours are the most promising arenas of outlier determination. Detection using statistical methods is defined in this model. In any data, the detection of outlier estimates important and critical information in wide application domains. The outlier detection has multiple formulations and they have been discovered in multiple disciplines like statistics, machine learning, data mining and information theory. The analysis of medical data making use of DM techniques is practically an untouched subject and is in need of some extra focus. In this study, the Pima data set was used in the simulation carried out by TANAGRA. A total of 193 outliers were detected for the statistics namely leverage, R-standard, R-student, DFFITS, Cook's D and covariance ratio. Its conclusions show us that exceptional behaviour of outliers facilitates the exploration of essential data which may be in hiding among the domains of the same. This assists the decision makers in giving better, dependable and competent healthcare services.

METHODOLOGY

In the domain of diabetes diagnosis, most of machine learning models concentrate on the learning of the Pima Indian Diabetes dataset in the UCI repository. Such a dataset has been largely utilized in ML experiments. It is presently accessible through the UCI repository of standard datasets. The National Institute of Diabetes, Digestive and Kidney Diseases has examined and studied this population constantly due to the rising diabetes patients. The Pima dataset consists of 768 data samples and it is used for the negative as well as positive effects of the diabetes disease. There are 8 risk attributes for diabetes namely Plasma glucose concentration, Diastolic blood pressure (mmHg), Triceps skin fold thickness (mm), 2-hour serum insulin (mu

Submitted : 24th August, 2019 | Accepted : 19th September, 2019 | Publication : 15th December, 2019

U/ms), Body mass index (weight in kg/(height in m))/2, Diabetes pedigrees function, Age (years). Training sets of 576 cases (378 non-diabetics, 198 diabetics) and 192 cases (122 non-diabetics, 70 diabetics) were created from the total number of cases. 268 diabetic patients (represented as "1") and remaining (represented as "0") are taken into consideration. A brief of Regression is presented below.

REGRESSION

A machine learning; data mining technique called Regression is mainly used for the fitting of an equation to a dataset. Two main linear regressions namely simple linear regression and multiple linear regressions (also called multivariate linear regression) exist. Simple linear regression deals with one dependent and one independent variable (outcome or response). On the other hand multiple linear regression deals with one dependent and two or more independent variables. Linear regression i.e. the simplest form of regression uses the straight line formula

(y = mx +b) to figure the suitable m and b values in order to predict y based on x. More complex models can be fit using progressive techniques that allow multiple input variables.

The models of regression are evaluated on several statistics forecasting the predicted values and the expected values and the difference between them. A regression project's historical data is stereotypically divided into two data sets; one for building the model and the other for testing it. Regression modeling has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug response modeling and environmental modeling. The anomalous observations are in general considered outliers or 'influential' data points by statisticians. However, they are considered outliers, high leverage points and influential observations in regression analysis. As we already know that an outlier is an anomaly in a data set, it can sometimes be doubted to have been created by some other mechanism. Statistical outlier detection techniques are essentially model-based techniques. In case of low probability of an instance being created by this model, the instance is considered an outlier. In the early 19th century, the outlier detection need was noticed by the statisticians. Outlying enabled them to perform accurate analysis on the data. This led to the notion of accommodation or removal of outliers in different statistical techniques. Regression model based on outlier detection techniques typically analyze the residuals obtained from the model fitting process to determine how outlying is an instance with respect to the fitted regression model and Grubbs' test is also called as ESD method. The initial stage will be to detect the deviation of the outlier from others and then to calculate the ratio as the difference between the outlier and the mean divided by the SD. Larger the value of Z, larger the deviation. Finally, the mean and SD are calculated from all the values including the outlier.

TREATMENT OF OUTLIERS

The above procedure can only help us determine doubtful points from the perspective of statistics. It however doesn't in anyway implying their elimination as that can be dangerous. In many cases, removing them may improve the regression "fit" but it may eliminate some key points from the data. Some standard methods to detect outliers are:

(i) Eyeball Method
(ii) Standardized or Studentized Residual Scores
(iii) Leveragability Statistics (Hat Values)
(iv) Distance D as in Cook's D

EXPERIMENTS AND RESULTS

Since the enactment of a multiple linear regression analysis for a large set of data would be immensely time consuming, the use of statistical analysis software for the quicker test performance. The results clearly calculate as below, also presented in figs 3.2 to 3.14:

(i) R^2 value
(ii) p-value

The coefficient of determination R^2 is an essential tool to assess the model fit. The regular R^2 always increases with increased number of factors while the adjusted R^2 considers the model complexity. A good model should maximise the adjusted R^2 i.e. a measure of the precision of well predicted future outcomes. Adjusted *R*2 is an *R*2 that adjusts for a number of explanatory terms in a model; which increases an improvement in the model occurs due to the introduction of a new term. There exists a chance of the adjusted *R*2 to be negative and hence would be less than or equal to *R*2. An **F-test** is any statistical test in which the test statistic has the F-distribution under the null hypothesis. The F-test in one-way analysis of variance is used to assess whether the expected values of a quantitative variable within several pre-defined groups, differ from each other. The alpha value arising from a test gives the **p-value. "Degrees of freedom"** is an integer value measuring the extent to which an experimental design imposes constraints upon the pattern of the mean values of data from various meaningful subsets of data. Lower p-value than the substantial level of test signifies the importance of the model. **Residual** is defined as the error predicted from the difference between the predicted value and the actual value. The kurtosis is observed to be of sub Gaussian type.

Regression Assessment Parameters Used data set: selected examples

RESULTS DATA SET SIZE: 768

TREE

PG < 144.5000
BMI < 28.8500 then **avg(CLASS) = 0.0701** (std-dev = 0.2561, with 157 examples [30.54%])
BMI >= 28.8500
AGE < 29.5000
PG < 127.5000 then **avg(CLASS) = 0.1495** (std-dev = 0.3583, with 107 examples [20.82%])
PG >= 127.5000 then **avg(CLASS) = 0.5882** (std-dev = 0.5073, with 17 examples [3.31%])
AGE >= 29.5000 then **avg(CLASS) = 0.4737** (std-dev = 0.5015, with 114 examples
[22.18%])
PG >= 144.5000 then **avg(CLASS) = 0.6975** (std-dev = 0.4613, with 119 examples [23.15%])

The dataset is split into growing and pruning sets by the regression algorithm. A two-step algorithm was used wherein a maximal tree fitting the possible growing set was built in the first step and nested sub-trees were tested as per the cost complexity principle. The optimal tree was selected on the pruning set and the simplest sub-tree with performance close to the optimal tree was selected on the growing set.

CONCLUSION

Routine health check-ups is a common practice among adults in most developed countries. Therefore, lesser expensive precautionary measures in case of detection of any disease in its early stages of development gives a patient a better chance at survival than detection of the same at a later stage. Clinical databases with patient information are essential to medical researchers and doctors. In fact, the study with medical data by using the DM techniques is virtually an unexplored frontier which needs extraordinary attention. It can be suggested that:

(i) The anomalous outlier conducts facilitate the survey of valuable information buried in their domain. This in turn assists the decision makers in their functioning.
(ii) The present experimental results can be used by the

medical doctors to sensibly predict tools from the vast medical database.

(iii) Some of the most promising areas would be patient symptoms, diagnoses and behaviours as well as the thorough understanding of their complex relationships.

For the experiment to be executed on outlier detection, five medical datasets namely viz., liver (345,7), Parkinson(198,24), Heart(270,14), Thyroid(216,6), Haberman (306,4) as instances and attributes are utilized respectively. From the present statistical analysis, it is found that 78 outliers for liver, 67 for Parkinson, 82 for heart, 110 for Thyroid and 61 for Haberman medical datasets are detected.

REFERENCES

1. Aboul Ella Hassanien and Jafar, M.H. Ali, "Rough Set Approach for Generation ofClassification Rules of Breast Cancer Data", Informatica, Vol. 15, No. 1, 23–38, 2004.
2. Aggarwal C.C. and Yu P., "Outlier detection for high dimensional data", Proceedings of ACM SIGMOD International Conference on Management of Data, 2001.
3. Ali K. M. and Pazzani M. J., "Error Reduction through Learning Multiple Descriptions", Journal of Machine Learning, 24:3, 173-202, 1996.
4. Alvin C. Rencher, "Methods of Multivariate Analysis", Wiley Interscience, second edition, 2002.
5. Amardeep Kaur, Dr. Bhatia M.P.S., Dr. Bhaskar S.M., "State Of the Art of OutlierDetection in Streaming Data", IADIS European Conference Data Ming, 2007.