Novel method in detecting outliers in medical database

Hiren R. Kavathiya
Assistant Professor
Department of CS & I.T.
Shree M.V. & Smt. N.Virani Science College,
Rajkot

Dr. G. C.Bhimani
Professor & Head,
Department of Statistics
Saurashtra University, Rajkot

Abstract:
In this paper, we used two models of outlier detection techniques in which the first model talks about Application of Data Mining Techniques for Outlier Mining in Medical Databases and the second model throws light on Outlier Mining in Medical Databases by UsingStatistical Methods. Both the models emphasizes on detecting the outliers in the medical databases by the way of mining through the entire database. First model makes use of the statistical analysis tools for the work and takes care of complicated issues in terms of patient symptoms, diagnoses and behaviors and hence they are said to be the most promising arenas of outlier determination. In second model outliers of 5 datasets; them being leverage, R-standard, R-student, DFFITS, Cook's D and covariance ratio are taken care off and explained.

Keyword: Data Mining, Outliers detection, Statistical analysis, Medical Databases

Introduction:

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.Distance based, density based and distribution based methods are the common outlier detection techniques. Outlier detection can be defined as the search of objects that do not follow valid rules for major section of data in databases. Its identification as being an outlier depends on aspects, mostly practical application based. An instance is the unruly flow of network packages. These unruly packages is determined by the system analysis log, chances of being an outlier as it might be a virus attack or any other kind of a security threat.The outlier detection problem and the classification problem are very similar. Most database objects analysed for the outlier detection problem aren't outliers and as in several cases, it's not a priori known what objects would be outliers.Outliers have several origins and their detection has become tougher with the constant growth of medical dataset. Most of the existing approaches to the problem of outlier detection in the literature have been based on density estimation methods, and in particular, on nearest-neighbour methods.On a general note, the outlier detection performance largely depends on the precise choice of the sampling distribution of the artificial examples. Datasets available at repositories help in judging the effectiveness of our approach and its results show the superiority of our methods over others. Regular health checkups are a common practice among adults of developed countries. They believe in the simple principle of "prevention is better than cure" and hence any early detection of disease gives the patient a chance to start the treatment at the right time. The exact healing of the patient may not be known but the treatment would provide the patient with life comfort. The blood count is the most important aspect of the health routine test and its results are the first to be looked at by the doctor to determine the patient's overall health since the standard range is common knowledge. Performing such a test has become easier overtime due to the automatic and semi-automatic machines available that perform the task for us. Although it is pretty accurate, it still needs to be checked by a physician since the normal range of blood count varies from individual to individual. The general conditions of the patient are influenced by several factors such as drugs, other treatments, sleep, etc. For such cases, the different values need to be recorded. 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.

Model 1:
Application of Data Mining Techniques for Outlier Mining in Medical Databases

The subject of outlier detection has been very crucial for data analysis and any complicated issues in terms of patient symptoms, diagnoses and behavioursare 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 U/ms), Body mass index (weight in kg/(height inm))/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.

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 ofthe precision of well predicted future outcomes. Adjusted $R2$ is an $R2$ 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 $R2$ to be negative and hence would be less than or equal to $R2$.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 α 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 subGuassian type. **Regression AssessmentParametersUsed data set:** selected examples
ResultsData set size: 768
Fig. 3.9revealsthe existence of internal inconsistencies of the data set clearly suggesting that the data set is corresponded to diabetic patients.

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.

Model 2: Outlier Mining in Medical Databases by Using Statistical Methods

The medical and public health domain's outlier detection works along patient records and is a very sensitive issue. We establish a total of 78, 67, 82, 78 and 69 outliers of 5 datasets; them being leverage, R-standard, R-student, DFFITS, Cook's D and covariance ratio. Below are some suggestions of the same.

(i) The anomalous outlier conducts facilitate the survey of valuable in formation 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.

Methodology

Outlier detection in the medical and public health domains typically work with patient records. Anomalous patient condition, instrumentation or recording errors, etc. are some reasons for the existence of outliers in the data. The detection of outliers is not only a sensitive problem but also requires a high level of accuracy. Different features like age, vlood group, weight etc. are certain characteristics of the medical data and might also have temporal or spatial aspects. Most of the current outlier detection schemes aim at detecting outlying records. Effective outlier detection necessitates a model creation representing the data precisely. Many techniques have been developed over the years. However, real-world data sets and environments present a range of difficulties that limit the effectiveness of these techniques. One of the issues is with the data attributes that may be heterogeneous mixture of nominal types (relatively small in number with partial ordering) or continuous types (numeric with total ordering). The assumption of the outliers being either continuous or categorical is made by many outlier detection techniques and having a different attribute type makes it difficult to find relations between them as well as to define distance or similarity metrics for such data points. It is common to see that many techniques homogenize the attributes by discretization or conversion of categorical attributes into continuous by the application of some arbitrary ordering. This can also lead to information loss and noise increase. Methods for detecting outliers based on the regression analysis are also classified among statistical methods. It is also important to know that the issue is formulated for examining the conditional probability distribution. Regression models whose basis is any outlier detection technique evaluate those residents obtained from the model fitting process in order to define the process of outlying an instance in reference to the fitted regression model. The initial stage is to detect the deviation of the outliers from others and then calculate the ration Z 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 values including the outlier

Results

Table 1: Data set description

SI. 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**

A very time consuming process would be to analyse the performance of a multiple regression on a large data set and for those reasons, statistical analysis software can be used. Its outcomes show:

(i) R^2 value

(ii) p-value

The results are given in the table

Table 2: Global results

Dataset	**Endogenous attribute**	**Sample**	R^2	**Adjusted-R^2**	**Sigma error**	**F-test**
Bupa liver	Selector	345	0.1336	0.1182	0.4641	8.6909
Parkinson	Status	198	0.4927	0.4278	0.3266	7.5945
Statlog Heart	Status	270	0.5452	0.5221	0.3441	23.6111
Thyroid	Class	216	0.4633	0.4504	0.5387	36.0869
Haberman	Status	306	0.0898	0.0808	0.4149	9.9395

In the above table, the quantities R^2 , Adjusted- R^2, F-test and degrees of freedomhave predefined meanings.

Table 3: Analysis of Variance

Dataset	**Regression**				**Residual**			**Total**	
	xSS	**d.f.**	**xMS**	**F**	**xSS**	**d.f.**	**xMS**	**xSS**	**d.f.**
Bupa liver	11.23	6	1.87	8.69	72.82	338	0.2155	84.05	344
Parkinson	17.82	22	0.81	7.59	18.35	172	0.1067	36.18	194
Statlog Heart	36.34	13	2.79	23.61	30.31	256	0.1184	66.66	269
Thyroid	52.36	5	10.47	36.08	60.65	209	0.2902	113.02	214
Haberman	5.13	3	1.711	9.93	51.12	305	0.1722	57.12	305

Conclusion

Following conclusions can be made from the results

(i) The anomalous outlier conducts facilitate the survey of valuable sinformation 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 61for 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 MultipleDescriptions”, Journal of Machine Learning, 24: 3, 173-202, 1996.
4) Alvin C. Rencher,“Methods of Multivariate Analysis”,WileyInterscience, 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.
6) **Cheick-OumarBagayoko,Jean-Charles Dufour, Saad Chaacho3, Omar Bouhaddou, Marius Fieschi**, “Open source challenges for hospital information system(HIS) in developing countries: a pilot project in Mali”, Bagayoko et al. BMC MedicalInformatics and Decision Making,10-22, 2010.
7) **Chien-Tsai Liu , Pei-Tun Yang , Yu-Ting Yeh, Bin-Long Wang,** “The impacts of smart cards on hospital information systems--an investigation of the first phase of the national health insurance smart card project in Taiwan”, International Journal of Medical Infromatics,173-81,2006.
8) **Ching Wei Wang,** “New Ensemble Machine Learning Method for Classification and prediction on Gene Expression Data”, Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, 2006.
9) **Chiu A. L and Fu A. W.**, “Enhancement on local outlier detection”, Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS’03), pg 298-307, 2003.
10) **Cover T.M and Hart P.E.**, “Nearest neighbors pattern classification,” IEEE Trans on Information Theory, vol. 13, pg 21-27, 1967.
11) **Cristina G. Dascălu, Corina Dima Cozma, Elena Carmen Cotrutz**, “Observations about the Principal Components Analysis and Data Clustering Techniques in the Study of Medical Data”, World Academy of Science, Engineering and Technology 17,pg 69-73, 2006.