Outlier Mining in Medical Databases: An Application of Data Mining in Health Care Management to Detect Abnormal Values Presented In Medical Databases

Varun Kumar
Institute of Technology & Management,
Gurgaon, Haryana, INDIA

Dharminder Kumar
Guru Jambheshwar University of Science & Technology,
Hisar, Haryana, INDIA

R.K. Singh
M.P. Bhoj (Open) University,
Bhopal,
Madhya Pradesh, INDIA

Abstract
Outliers in medical databases can be caused by measurement errors or may be the result of inherent data variability. The abnormal value of *mitoses*, for instance, could lead to the diagnosis of *malignant cancer* or it might just be due to human mistake or execution error. In this paper, we make use of a large database, namely, *Wisconsin Breast Cancer Database* containing 10 attributes and 699 instances to detect outliers. Many data mining algorithms try to minimize the influence of outliers which could result in the loss of important hidden information since "one person's noise could be another person's signal". In particular, we used *TANAGRA* (A Data Mining Tool) to detect outliers from Breast Cancer Database and analyzed them for knowledge discovery. The results of the experiment show that outlier mining i.e. outlier detection & analysis have a great potential to find useful information from health care databases which consequently helps decision makers to automate & quicken the process of decision making in clinical diagnosis as well as other domains of health care management.

Key words:
Health Care, Breast Cancer, Data Mining, Outlier Mining, TANAGRA

1. Introduction

The healthcare environment is generally perceived as being 'rich in information' yet having 'knowledge poor' [10]. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Valuable knowledge can be discovered from application of data mining techniques in healthcare system [11]. The healthcare administrators to improve the quality of service can use the discovered knowledge. Very often, there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers [14]. Outliers can be caused by measurement or execution error. For example, the display of a person's age as -999 could be caused by a program default setting of an unrecorded age. Alternatively, outliers may be the results of inherent data variability such as abnormal value of mitoses say 10 may signifies as one of the main cause of suffering a patient with malignant tumor.

The database used for this experiment is about the diagnosis of breast cancer. The body is made up of many types of cells. Normally, cells grow and divide to produce more cells only when the body needs them. This orderly process helps keep the body healthy. Sometimes, however, cells keep dividing when new cells are not needed. These extra cells form a mass of tissue, called a growth or tumor [4]. Tumors can be *benign* or *malignant*.

Benign tumors are not cancer. They can usually be removed, and in most cases, they do not come back. Cells from benign tumors do not spread to other parts of the body. Most important, benign breast tumors are not a threat to life. Malignant tumors are cancer. Cells in these tumors are abnormal. They divide without control or order, and they can invade and damage nearby tissues and organs. That is how cancer spreads from the original (primary) cancer site to form new tumors in other organs. The spread of cancer is called metastasis.

When cancer arises in breast tissue and spreads (metastasizes) outside the breast, cancer cells are often found in the lymph nodes under the arm (axillary lymph nodes). The disease is called metastatic breast cancer [4].

1.1 Motivation

According to American Cancer Society in 2004 (the latest year for which figures are available), approximately 2.4 million women living in the U.S. had a history of breast cancer. In 2007, approximately 40,460 women are expected to die from breast cancer. According to Times of India report last year in India, nearly one lakh women died from breast cancer. By 2015, there will be approximately 2.5 lakhs new cases in India. According to World Cancer Report in the year 2000, malignant tumors were

Manuscript received August 5, 2008.
Manuscript revised August 20, 2008.

responsible for 12 per cent of the nearly 56 million deaths worldwide from all causes. Cancer rates could further increase by 50% to 15 million new cases in the year 2020. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care. The research studies on health care management aim both to control the increasing costs and to increase the accessibility level for health care services [13]. In the global scenario of health care there exists many loopholes as the medical data is mostly, not completely accurate. There could be improvement of our health care system with the use of Data Mining. [15].

1.2 Organization of the Paper

The paper is organized as follows: Section 2 defines problem statement and related work in this area. Section 3 describes the proposed method to detect outliers breast cancer database and to analyze them using outlier mining. Section 4 presents the snap shots of Wisconsin Breast Cancer Database containing 10 attributes and 699 instances used to perform this experiment. Experimental results and their analyses are presented in Section 5 and finally, Section 6 concludes the paper and points out some potential future work.

2. Problem Statement and Related Works

We are applying outlier mining to breast cancer health care databases so that useful information in the form of novel & hidden knowledge patterns can be generated to improve public health. The exact causes of breast cancer are not known. However, studies show that the risk of breast cancer increases as a woman gets older. With increase awareness in outlier detection in data mining literature, more concrete meanings of outliers can be defined for specific domains [9]. Thus the problem in particular is to detect outliers from breast cancer health care databases proceeded by an analysis work to prevent any medical error or to improve the detection of this disease in women.

Knorr's [7] work on mining distance based outlier in large databases shows a great potential to have same kind of applications in medical databases. The statistical approach and discordancy tests are described in V. Barnett and Lewis [14]. Nada Lavrac [13] worked on selected techniques for data mining in medicine. Density based outlier detection was first proposed by Breunig et al. [12]. It relies on the local outlier factor (LOF) of each point, which depends upon the local density of its neighborhood.

3. Proposed Method

Outlier mining can be described as follows: Given a set of n data points or objects, and k, the expected number of outliers, find the top k objects which are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data. The outlier mining problem can be viewed as two subproblems:

(1) Define what data can be considered as inconsistent in a given data set; and

(2) Find an efficient method to mine the outliers so defined.

The problem of defining outliers is nontrivial. If a regression model is used for data modeling, analysis of the residuals can give a good estimation for data "extremeness" [7]. The task becomes tricky, however, when finding outliers in time series data as they may be hidden in trend, seasonal, or other cyclic changes. When multidimensional data are analyzed, not any particular one, but rather, a combination of dimension values may be extreme. For non-numeric (i.e., categorical data), the definition of outliers requires special consideration [1].

In this paper we make a use of Wisconsin Breast Cancer Database which is already available in the desired format of TANAGRA i.e. in .txt file. After loading the desired database in TANAGRA we perform outlier mining using Outlier Detection component of TANAGRA as shown in *Table 1*.

Table 1: Components of TANAGRA used in this experiment

Tab	***Operator(Component)***	***Comment***
Feature selection	Define status	Specify the attributes to use
Statistics	Univariate Outlier Detection	To perform outlier mining
Instance Selection	Rule based selection	To generate rules for discovery knowledge
Data Visualization	Scatterplot with label	To visualize the relationship between two variables

After detecting outliers from the breast cancer medical database we analyzed them with the help of other components of TANAGRA. Outliers in medical databases can be caused by measurement errors or may be the result of inherent data variability. One person's noise could be another person's signal, in other words, the outliers themselves may be of particular interest, such as abnormal value of an attribute may be observed as strong symptoms of a particular disease. We analyzed the data from both view points.

4. Database used in Experiment

The Wisconsin Breast Cancer Database is originally available on UCI Machine Learning Repository website http://archive.ics.uci.edu:80/ml/datasets.html in Excel Format i.e. .xls file. In order to perform experiment using TANAGRA, the file format for breast cancer database has been changed to .txt file and a sample of records are as shown in *Table 2*. This dataset is containing 10 attributes and 699 instances. The complete description of the of attribute value are presented in *Table 3*.

Table 2: Sample of instances form Wisconsin Breast Cancer Dataset

clump	*ucellsize*	*ucellshape*	*mgadhesion*	*sepics*	*bnuclei*	*bchromatin*	*normnucl*	*mitoses*	*class*
4	2	2	1	2	1	2	1	1	Begnin
1	1	1	1	2	1	2	1	1	Begnin
2	1	1	1	2	1	2	1	1	Begnin
10	6	6	2	4	10	9	7	1	Malignant
4	1	1	1	2	1	2	1	1	Begnin
1	1	1	1	2	1	1	1	1	Begnin
1	1	1	1	2	1	2	1	1	Begnin
5	1	1	1	2	1	2	1	1	Begnin
3	1	1	1	2	1	2	1	1	Begnin
1	1	1	1	2	4	2	1	1	Begnin

Table 3: Complete description of variables

Variable/Attributes	*Category*	*Possible Values*
clump	Continue	1 – 10
ucellsize	Continue	1 – 10
ucellshape	Continue	1 – 10
mgadhesion	Continue	1 – 10
sepics	Continue	1 – 10
bnuclei	Continue	1 – 10
bchromatin	Continue	1 – 10
normnuct	Continue	1 – 10
mitoses	Continue	1 – 10
class	Discrete	Two, Begnin & Malignant

5. Experimental Results & Analysis

In order to perform this experiment breast dataset was given to Tanagra and outlier detection has been performed using Univariate Outlier Detection component. Lower Bound (L.B.) & Upper Bound (U.B.) of each attribute have been calculated using following parameters:

$$p - value\_for\_Grubb's\_test = 0.05 \quad (1)$$

$$multiple\_of\_sigma = 3 \quad (2)$$

Following observations (ref. *Table 4*) have been providing using above mentioned parameters (*Eq. 1* & *Eq. 2*) to *Univariate Outlier Detection* component of TANAGRA:

Table 4: L.B. & U.B. of each attributes with outliers detected

Attributes	*L.B.*	*U.B.*	*Outliers Detected*
clump	-4.0295	12.8650	0
ucellsize	-6.0199	12.2889	0
ucellshape	-5.7083	12.1223	0
mgadhesion	-5.7593	11.3730	0
sepics	-3.4259	9.8589	31
bnuclei	-7.2524	14.3626	0
bchromatin	-3.8773	10.7529	0
normnuct	-6.2939	12.0279	0
mitoses	-3.5558	6.7346	31

Now we have L.B. and U.B. of each variable and it is clear that there is no abnormal value or outlier detected with reference to *lump, ucellsize, ucellshape, mgadhesion, bnuclei, bchromatin, normnuct* as all the values of these variables presented in the database falls between the respective ranges as shown in *Table 4*. Only two variables namely, *sepics* and *mitoses* show the presence of abnormal or outliers values as 31 values of each variables are beyond the ranges mentioned in *Table 4*. A pictorial representation of outlier detected with reference to sepics and mitoses can also be observed through *Fig. 1* and *Fig. 2*.

A total of 51 outliers have been detected from the breast cancer database due to the abnormal values of sepics and mitoses. The details of these outliers are as given below is according to the following format:

[Patient's ID Number, No. of Outlier Detected, Concerned Variables]

[20, 2, sepics; mitoses], [27, 1, sepics], [82, 2, sepics; mitoses], [98, 1, mitoses], [99, 1, sepics], [102, 2, sepics; mitoses], [109, 1, sepics] [116, 2, sepics; mitoses], [124, 1, mitoses], [127, 1, mitoses], [140, 1, mitoses], [156, 2, sepics; mitoses], [181, 1, mitoses], [183, 1, sepics] [189, 1, sepics] [196, 1, mitoses], [225, 1, sepics] [227, 1, sepics] [234, 1, mitoses], [250, 1, mitoses], [252, 1, mitoses], [255, 1, mitoses], [263, 2, sepics; mitoses], [302, 1, sepics] [303, 1, sepics] [333, 1, sepics] [352, 2, sepics; mitoses], [353, 1, mitoses], [381, 1, mitoses], [409, 1, mitoses], [412, 1, sepics] [428, 2, sepics; mitoses], [448, 1, sepics] [457, 1, mitoses], [466, 1, sepics] [507, 1, mitoses],

[524, 1, sepics] [551, 1, sepics] [561, 1, mitoses], [587, 1, mitoses], [588, 1, mitoses], [589, 2, sepics; mitoses], [594, 1, sepics] [603, 2, sepics; mitoses] [614, 1, sepics] [615, 1, sepics] [616, 2, sepics; mitoses], [632, 1, mitoses], [653, 1, mitoses], [697, 1, sepics] [698, 1, sepics]

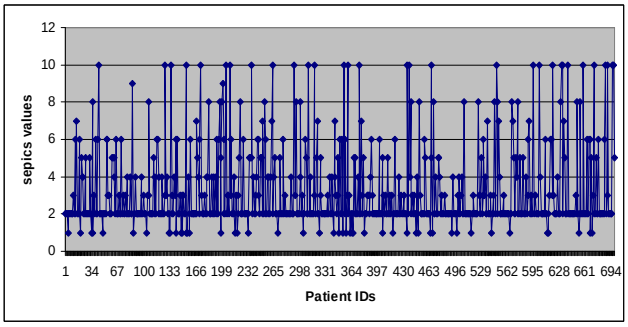

Fig. 1: Outliers detected with reference to sepics abnormal values

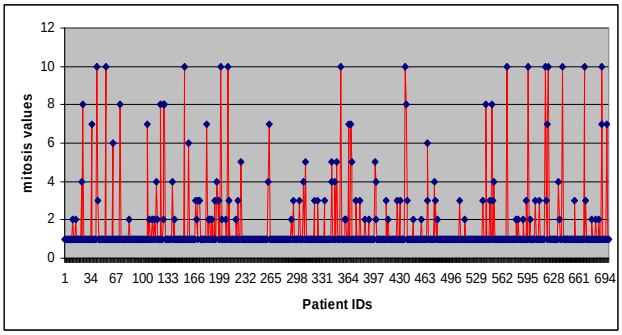

Fig. 2: Outliers detected with reference to mitoses abnormal values

Let's take a sample from the 51 outliers detected above i.e. *[82, 2, sepics; mitoses];* it means two outliers have been detected in the record of Patient ID 82, with reference to sepics and mitoses variables. By closely examined the tuple we have following observations:

- The value of *sepics* variable is 10 which is an abnormal value or outlier in a sense that it lies beyond the range (*ref. Table 4*).
- The value of *mitoses* variable is 10 which is an abnormal value or outlier in a sense that it lies beyond the range (*ref. Table 4*).
- The values of *sepics* and *mitoses* variables are very high in comparison with the values of these variables with respect to most of the patient's records.
- The patient with id no. 82 is classified with malignant class of tumor. Malignant tumors are cancer. Cells in these tumors are abnormal as shown in *Fig. 3*.

	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	class
76	4	1	1	1	2	1	3	1	1	begnin
77	1	1	1	3	2	1	1	1	1	begnin
78	8	10	10	1	3	6	3	9	1	malignan
79	5	1	1	1	2	1	3	1	1	begnin
80	3	1	1	2	2	1	1	1	1	begnin
81	6	5	5	8	4	10	3	4	1	malignan
82	9	6	9	2	10	6	2	9	10	malignan
83	2	1	1	1	2	1	3	1	1	begnin
84	1	1	1	1	1	1	3	1	1	begnin
85	5	1	1	1	2	1	1	1	1	begnin
86	1	1	1	2	1	1	1	1	1	begnin
87	7	1	2	3	2	1	2	1	1	begnin
88	3	1	1	1	2	1	1	1	1	begnin

Figure 3: Patient ID 82 with entire attribute values *[82, 2, sepics; mitoses]*

Now the question arises here is that how we can utilize this piece of knowledge generated with the help of outlier mining? In contrast to traditional data mining task that aims to find the general pattern applicable to the majority of data, outlier detection targets the finding of the rare data whose behavior is very exceptional when compared with the rest large amount of data.

As observed in the above experiment the abnormal value of sepics and mitoses in prima-facie may be caused by measurement or execution error but considering attribute details (ref. *Table 3*) these values are within range. The other side of the coin is the outliers themselves may be of particular interest, and may signify some hidden & implicit information. One important observation with patient id 82 is, two outliers have been detected with reference to sepics and mitoses variables and the patient is classified with malignant class of tumor.

If we apply same outlier values as a rule (shown in *Fig. 4*) to select instance from the entire database followed by identifying the class of other patients; we can observe that all the patients (instances) are suffering from malignant tumor as shown in *Fig. 5*.

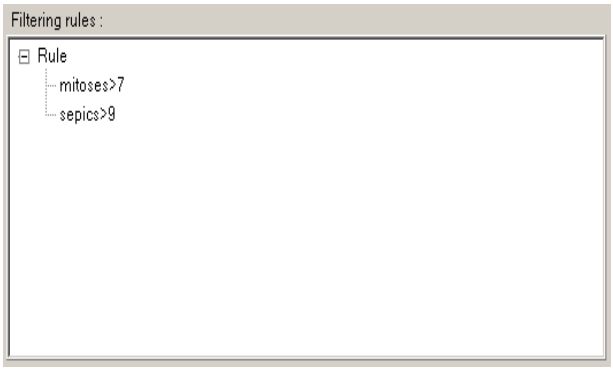

Figure 4: Detected outlier values as a rule to select instances from dataset

	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	class
82	9	6	9	2	10	6	2	9	10	malignant
102	5	10	10	10	10	2	10	10	10	malignant
116	8	10	10	7	10	10	7	3	8	malignant
156	6	10	10	10	10	10	8	10	10	malignant
263	9	10	10	10	10	5	10	10	10	malignant
352	10	10	10	10	10	1	8	8	8	malignant
428	5	10	6	1	10	4	4	10	10	malignant
589	6	10	2	8	10	2	7	8	10	malignant
616	10	10	10	10	10	10	4	10	10	malignant

Figure 5: Instances selected by TANAGRA based on rule mentioned in *Fig. 4*

In order to further prove our findings we used scatterplot component of TANAGRA to visualize the relationship between two variables i.e. sepics and mitoses as shown in *Fig. 6*. The outlier detected with reference to sepics and mitoses shows the tendency of patients to be classified as malignant class of tumor.

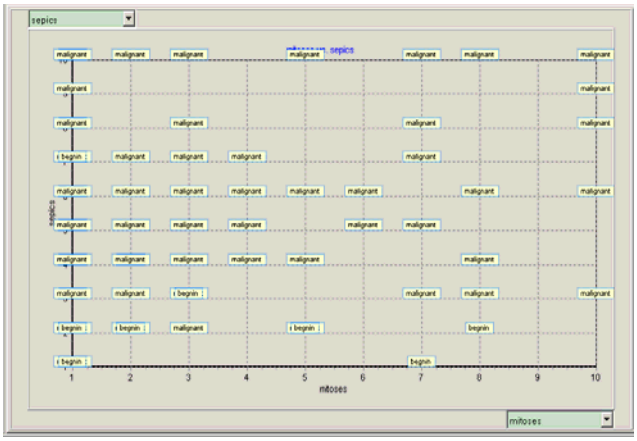

Figure 6: Graph visualizing relationship between two variables sepics and mitoses

6. Conclusion and Future Work

Today's clinical databases store detailed information about patient diagnoses, lab test results and details from patient treatments, a virtual gold mine of information for medical researchers. Utilizing data mining techniques with medical treatment data is a virtually unexplored frontier. Studying the extraordinary behavior of outliers helps uncovering the valuable knowledge hidden behind them and aiding the decision makers to improve the health care services.

The presented experiment give medical doctors a tool to help them quickly make sense of vast clinical databases. Understanding the complex relationships that occur among patient symptoms, diagnoses and behavior is one of the most promising areas of outlier mining. The Wisconsin Breast Cancer Database containing 10 attributes and 699 instances wherein all attributes are considered as categorical and the definition of outliers requires special considerations with categorical variables. In future work, we will study about applying comparative outlier mining in medical analysis for finding unusual responses to various medical treatments.

7. Acknowledgements

This research work is supported by the UGC Grant under Major Research Project Scheme (F. No.: 33-60/2007 (SR) dated: FEB 28, 2008).

8. References

[1] A. Agresti. An Introduction to Categorical Data Analysis. John Wiley & Sons, 1996.

[2] B. Chandrasekaran, F. Gomez, S. Mittal, J. Smith, “An Approach to Medical Diagnosis Based on Conceptual Structures,” Proceedings of the 6th International Joint Conference on Artificial Intelligence, Tokyo, Japan, 1979.

[3] David Riaño, Susana Prado (2000), “A data-mining alternative to model hospital operations: clinical costs and predictions”, Lecture Notes in Computer Science 1933, pp. 293-299

[4] D. Dearnaley et al., Handbook of adult cancer chemotherapy schedules, The Medicine Group (Education) Ltd., Oxfordshire, 1995.

[5] Desouza, K.C. (2001) Artificial intelligence for healthcare management In Proceedings of the First International Conference on Management of Healthcare and Medical Technology Enschede, Netherlands: Institute for Healthcare Technology Management.

[6] Dick, R.S., and Steen, E.B., “The Computer-Based Patient Record”. 1991, Washington, D.C.: Institute of Medicine. National Academy Press.

[7] E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. In Proc. 1998 Int. Conf. Very Large Data Bases, pages 392 - 403, New York, NY, August 1998.

[8] G. Piatetsky-Shapiro and W. J. Frawley. “Knowledge Discovery in Databases”. AAAI/MIT Press, 1991.

[9] K. Cios, W. Pedrycz, and R. Swiniarski. Data Mining Methods for Knowledge Discovery. Boston: Kluwer Academic Publishers, 1998

[10] Lincoln, T. & Builder, C. (1999) Global healthcare and the flux of technology. International Journal of Medical Informatics, 53. 213-224.

[11] Kristin B. DeGruy, Healthcare Applications of Knowledge Discovery in Databases, JOURNAL OF HEALTHCARE INFORMATION MANAGEMENT®, vol. 14, no. 2, Summer 2000.

[12] M.M. Breunig, H.P. Kriegle, R.T. Ng, J. Sander. LOF: Identifying Density based Local Outliers. In: proc. Of SIGMOD’00, pp. 93-104, 2000.

[13] Nada Lavrac (1999), “Selected techniques for data mining in medicine”, Artificial Intelligence in Medicine 16 3–23.

[14] V. Barnett and T. Lewis. Outliers in Statistical Data. John Wiley & Sons, 1994.

[15] Wasan Siri Krishan , Harleen Kaur, “Empirical Study on Application of Data Mining Techniques in Healthcare”, Journal of Computer Science 2(2): 194-200, 2006.
[16] Wennberg J, Cooper MM, editors. The Dartmouth atlas of medical care in the United States: a report on the Medicare program. Chicago, IL: AHA Press; 1999.

Varun Kumar received his M.Tech. in Information Technology and M.Phil. in computer Science. He is recipient of Gold Medal at his Master’s degree. Currently he is pursuing Ph.D. in Computer Science. He is having more than 10 years of teaching and research experience. He has published more than 20 research papers in Journals/Conferences/Seminars at national/International levels. His area of interest includes Intelligent Systems, Data Mining, Soft Computing, and Artificial Intelligence. He is now with department of Computer Science & Engineering and Information Technology, Institute of Technology & Management, Gurgaon, Haryana, INDIA.

Dharminder Kumar received his Ph.D. in the area of Computer Science in Computer Networks. He is recipient of Gold Medal at his Master’s degree. Currently he is heading the Faculty of Engineering and Technology as Dean and also the department of Computer Science & Engineering as Chairman. He has guided five students in Ph.D., about 40 at M.Tech. level and six students are currently under his supervision in pursuing their Ph.D. in Computer Science and Engineering. He has published more than 40 research papers in Journals/Conferences/Seminars at national/International levels. He is having more than 20 years of teaching and research experience. His area of interest includes Data Mining and Computer & Communication Networks. He is now with department of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology, Hisar, Haryana, INDIA.