International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 5 (2015) pp. 12627-12636

http://www.ripublication.com

Anomaly-Detection In Diabetes Using SVM

Jabez J
Research Scholar, Sathyabama University
Chennai., Tamil Nadu, India.
Dr.B.Muthu Kumar
Professor, Sathyabama University
Chennai, Tamil Nadu, India.

Abstract

Gradual multiplication of data over sectors provoked the discomfort in detection of abnormality or anomaly in such immense data, consequently the initiation of miscellaneous algorithms for such consequences had been generated. In this paper we have chosen One-Class Support Vector Machine for anomaly detection in diabetes, as it has the classification mining without target to detect the abnormality. The multitudinous factors on examination of the algorithm are studied and analyzed. The results of the algorithm were examined and analyzed for efficacy in progression of the algorithm.

Introduction

Anomaly detection predominantly identifies the cases which show abnormality that are unusually present within the data having seemingly the homogeneous characteristics. In the case of fraud detection, network intrusion, medical stream and other rare events that have great significance but are hard to find, anomaly detection becomes a predominant tool in such cases.

With the help of Anomaly detection technique the following such problems could also be with held:

- Flagging of suspicious activities in data compiled by the law of enforcement agency about illegal activities, without legitimate activities is one of the sectors where anomaly detection is significantly major. As know that the law enforcement data is all of one class. There are no counter-examples for appropriate reasoning.
- Identification of fraudulent claims in millions of insurance claims processed by insurance agencies, keeping in mind the number of fraudulent in this circumstance being very small. In such cases there are a few counter-examples such as outliers.

Diabetes has been the most common disorder for almost every individual in the contemporary days, even small kids are prone to have diabetes either by hereditary or by consumption of high glucose where circumstances show the inadequate production of insulin in such situations. Detection of abnormality in such case where dataset number is enormous in size is considerably tough. Incorporation of anomaly detection algorithms in this sector would comfort the examiners in much better way. The dataset utilized for evaluation of the algorithms is the UCI's diabetes; the dataset description is given in section 3. The unspecified special events during the examination, under the observation circumstances are to be highlighted during the test results analysis.

Error rates of the approaches during the analysis of the tested results are evaluated on the experimental results of each approach. Each evaluation shows many variations during the study of the techniques individually; As a result although researchers have conducted experiments to estimate the performance of the techniques with the computational results, comparison of the resultants was unsustainable.

Related Works

A study was made on various outlying approaches in the ongoing works for anomaly detection out of which a few are listed below.

H. Jair Escalante [1] made a comparative study of outlier detection algorithms; he presented an overall overview on outlier detection methods and experimental results of 6 implemented methods. He even applied these methods for the prediction of stellar populations' parameters as well as on machine learning benchmark data, inserting artificial noise and outliers, using kernel principal component analysis in order to reduce the dimensionality of the spectral data. Experiments on noisy and noiseless data were also performed.

Another comparison study algorithms for keystroke dynamics was made by Kevin S. Killourhy and Roy A. Maxion [2], this keystroke dynamics is the analysis of typing rhythms to discriminate among users which was proposed for impostors detection, for this task many algorithms were proposed, so to measure the performance, they collected a keystroke-dynamics data set, to develop a repeatable evaluation procedure and compare the results soundly, their collection consisted of 51 subjects typing 400 passwords each, and we implemented and evaluated 14 detectors from the keystroke dynamics and pattern-recognition literature. On resultant, the three top-performing detectors achieve equal error rates between 9.6% and 10.2%. The result along with the shared data and evaluation methodology constitutes a benchmark for comparing detectors and measuring progress.

Ingo Steinwart et al [3] introduced two new algorithms; a global variant of the cluster-based local outlier factor (CBLOF) which compensated the shortcomings of the actual method and the local density cluster-based outlier factor (LDCOF) which took the local variances of clusters into account. The performance was evaluated using real world datasets from the well know UCI machine learning repository. The strengths and weaknesses of the algorithms individually were revealed on resultant

and showed that our proposed clustering based algorithms out-perform CBLOF significantly.

A novel method to improve the performance of current AD algorithms was presented by Mohsen Zare Baghbidi et al [4]. Their proposed method calculated the Discrete Wavelet Transform (DWT) of every pixel vector of image using Daubechies4 wavelet; then, four bands of "Wavelet transform" matrix is performed by AD algorithm which are the approximation of original image. His research even included Local RX, DWRX and DWEST which are some of the benchmark AD algorithms for implementing on Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) hyperspectral datasets. Runtime of proposed method was significantly improved which is demonstrated by their experimental results. The accuracy of AD algorithms is also improved because of DWT's power in extracting approximation coefficients of signal, which contain the main behavior of signal and abandon the redundant information in hyperspectral image data.

Tarem Ahmed et al [5] used two different datasets, pictures of a highway in Quebec taken by a network of webcams and IP traffic statistics from the Abilene network, as examples in demonstrating the applicability of two machine learning algorithms to network anomaly detection. They then investigated the use of the block-based One-Class Neighbour Machine and the recursive Kernel-based Online Anomaly Detection algorithms.

Gowri, S and Anandha Mala, G.S [11] has proposed a system for retrieval of information of the purpose of forensic analysis in order to reduce to the effort of the analyzers for the detection of anomalies through the text retrieved appropriately from the massive data collected from the communication network channel, such as SMS, chat rooms, mails and so on. This is also one of the ways which comforts the users to analyze the data appropriately.

Data set Description

A. Data Set Information:

Two different sources were used for obtaining the patients records; one is an automatic electronic recording device and the other is manual recording in paper. An internal clock for timestamp of events was set in the automatic device, whereas only "logical time" slots such as breakfast, lunch, dinner and bedtime where only taken in the case of paper records. Fixed timings were assigned to breakfast, lunch, dinner, and bedtime as 08:00, 12:00, 18:00 and 22:00 respectively for paper recording purpose. Hence paper records have fictitious uniform recording times whereas electronic records have more realistic time stamps [6].

B. Attribute Information:

Diabetes files consist of four fields per record. Each field is separated by a tab and each record is separated by a newline.

File Names and format:

(1) Date in MM-DD-YYYY format
(2) Time in XX:YY format
(3) Code
(4) Value

	A	B	C	D	E
1	10/10/1989	8:00	58	149	
2	10/10/1989	8:00	33	10	
3	10/10/1989	12:00	60	116	
4	10/10/1989	12:00	33	4	
5	10/10/1989	18:00	62	304	
6	10/10/1989	18:00	33	10	
7	10/10/1989	22:00	48	63	
8	10/10/1989	22:00	33	14	
9	10/11/1989	8:00	58	171	
10	10/11/1989	8:00	33	10	
11	10/11/1989	12:00	60	148	
12	10/11/1989	12:00	33	4	
13	10/11/1989	18:00	62	115	
14	10/11/1989	18:00	33	10	
15	10/11/1989	22:00	48	130	

Figure 1: The input data set values in .csv format type

The Code field is deciphered as follows:

33 = Regular insulin dose
34 = NPH insulin dose
35 = UltraLente insulin dose
48 = Unspecified blood glucose measurement
57 = Unspecified blood glucose measurement
58 = Pre-breakfast blood glucose measurement
59 = Post-breakfast blood glucose measurement
60 = Pre-lunch blood glucose measurement
61 = Post-lunch blood glucose measurement
62 = Pre-supper blood glucose measurement
63 = Post-supper blood glucose measurement
64 = Pre-snack blood glucose measurement
65 = Hypoglycemic symptoms
66 = Typical meal ingestion
67 = More-than-usual meal ingestion
68 = Less-than-usual meal ingestion
69 = Typical exercise activity
70 = More-than-usual exercise activity
71 = Less-than-usual exercise activity
72 = Unspecified special event

Anomaly Detection Models Description

Anomaly detection is basically one form of classification like clustering, there are many types of classifications in data mining one of which is anomaly detection which would insist on abnormality check on more specific way.

There are various types of anomaly detection techniques based on the data.

Few types are: One-Class Classification, Single-Class Data and Finding Outliers

A. One-Class Classification

Anomaly detection is identified as one-class classification mainly due to representation of the input data i.e., the training data set is to be represented as just one class without any other fields to be taken into consideration for detection of anomaly. The main output of the anomaly detection model is to predict whether the data point is typical for a given distribution or variation of the point is detected. An atypical data point of the dataset might be either an outlier or an example of previously unseen class.

On testing phase of any anomaly detection algorithm the must be trained on input type of both examples and counter examples so as to make the designed model distinguish between normality and abnormality. For example a model designed for the purpose of the predicting the side effects of any random medication; in this case the anomaly detection model must be trained over the data that includes wide range of responses to the medication.

The description of typical cases in the training data is profiled and developed by the one-class classifiers. Any data point which show a deviation from the actual profile is identified to an anomaly. Basically the one-class classifiers are mostly identified as the positive security models due to the fact that they seek to identify the positive behaviors by assuming the rest as bad behavior orientated elements.

B. Single-Class Data

The classification of all the class is same in this single-class classification. Counter-examples: to collect the other class instances, it is hard to specify and expensive as well. For example it may be easy to classify a document under a given topic for a teat document classification. As the documents are not limited there aren't any counter-examples under this classification. Here the unusual instances are found in particular document type as the size of data to be classified is large. O Mazhelis [7] had made an analysis of Suitability in the Context of Mobile-Masquerader Detection which utilizes the One-Class Classification to employ the training data.

C. Finding Outliers

Outliers fall outside the distribution that is considered normal for the data so they are meant to be treated as unusual. For instance, Census Data [9] might show median household income as $70,000 and a average household income of $80,000, but a few households income might be of $200,000. Such cases would probably be identified as outliers. There are plenty of ways for outlier detection; Markus M. Breunig et al [8] has proposed a method named LOF which is used for the purpose of outlier detection based on the density.

The distance from the center of a normal distribution indicates how typical a given point is with respect to the distribution of the data. Each case can be ranked according to the probability that it is either typical or atypical.

The presence of outliers can have a deleterious effect on many forms of data mining. Anomaly detection can be used to identify outliers before mining the data.

For this diabetes analysis classification type chosen was the one-class classification due to 2 the presence of 2 independent attribute and 2 timestamp based attribute which are set under regulatory basis.

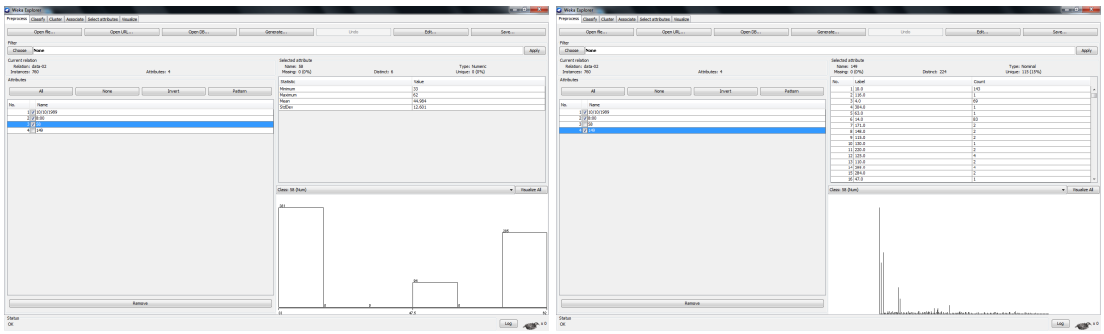
Results and Discussions

On use of SVM for anomaly detection only classification mining function is availed but no target on specific. One-class SVM models, when applied, produces a prediction and a probability for each case in the scoring data. If the prediction is 1, the case is considered typical. If the prediction is 0, the case is considered anomalous. This behavior reflects the fact that the model is trained with normal data.

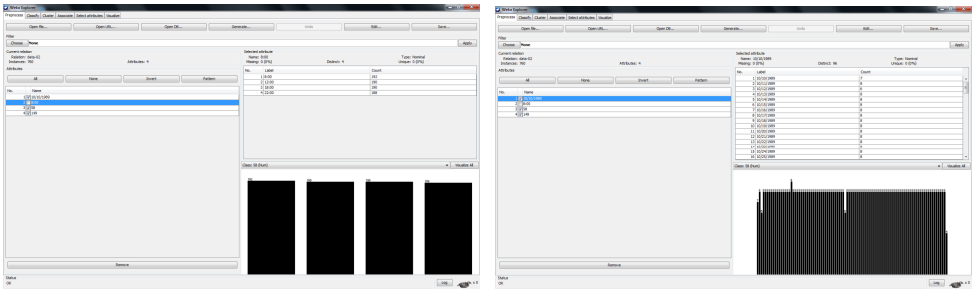
You can specify the percentage of the data that you expect to be anomalous with the SVMS_OUTLIER_RATE build setting. If you have some knowledge that the number of suspicious cases is a certain percentage of your population, then you can set the outlier rate to that percentage. The model will identify approximately that many rare cases when applied to the general population. The default is 10%, which is probably high for many anomaly detection problems.

This One-Class SVM model was implemented in the Weka tool [10] loading the algorithm as .jar file format and the screenshots of the derived classification technique are described in brief.

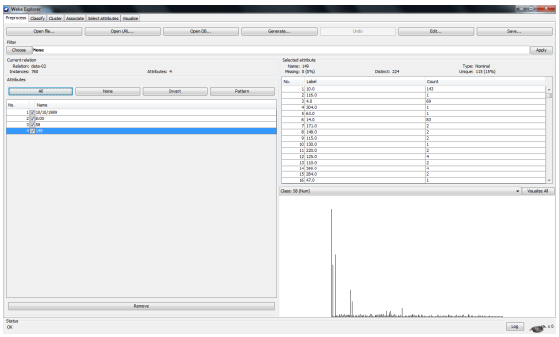
Initially the data which has to be classified has to be processed, i.e., it has to be distributed forming an uniformity over the enormously large data. For processing the values are made to be distributed and they are then plotted on a 2D graph which would be visualized with the default bar graph representation. The Data processing graphical representation is shown in the figure 2 below.



(a) (b)



(c) (d)



(e)

Figure 2: (a) Data-Time and Code, (b) Date-Time and Value, (c) Date, Code and Value, (d) Time, Code and Value, (e) Date, Time, Code and Value

The Layout formed for this anomaly detection model has the CSV loader, Attributes detector, SVM, Data visualization, Graph generator and a CSV server. The layout designed is shown in figure 3 below.

The CSV loader loads the input data after which the attributes are identified by the attributes detector, once the attributes are identified they are forward to the SVM processor to process the data and classify the data. The Classifications are then represented in both normal data matrix view as well as graphical view. After the classification the CSV loader saves the classified.

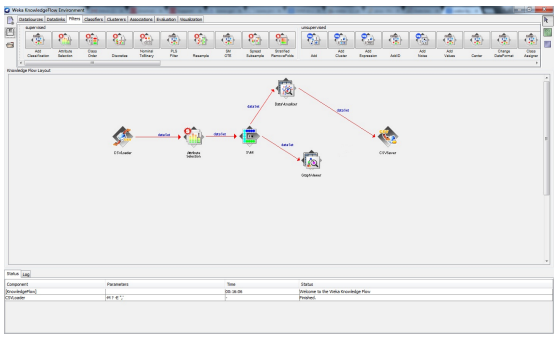


Figure 3: Layout of the Anomaly detection model

For any classification to be made the detail of the input data is to be know and defined in a clear format in-order to make the processing of any layout or algorithm perform much better than knowledge less processing, the figure 4 and 5 represents the classification formed.

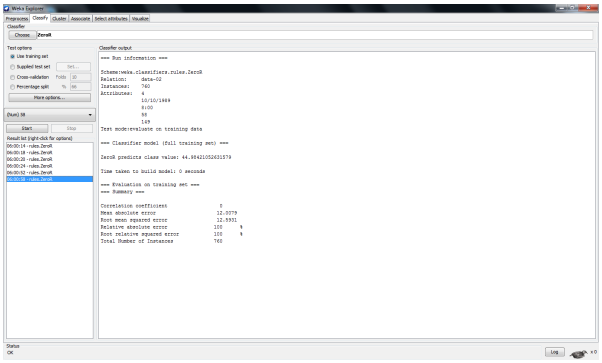
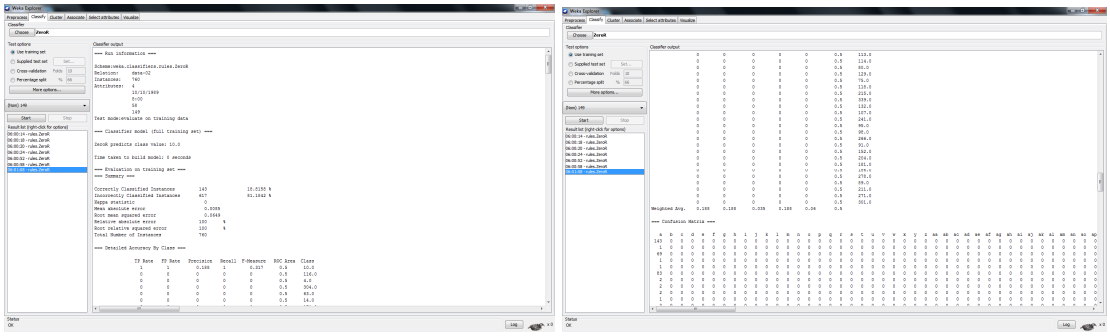


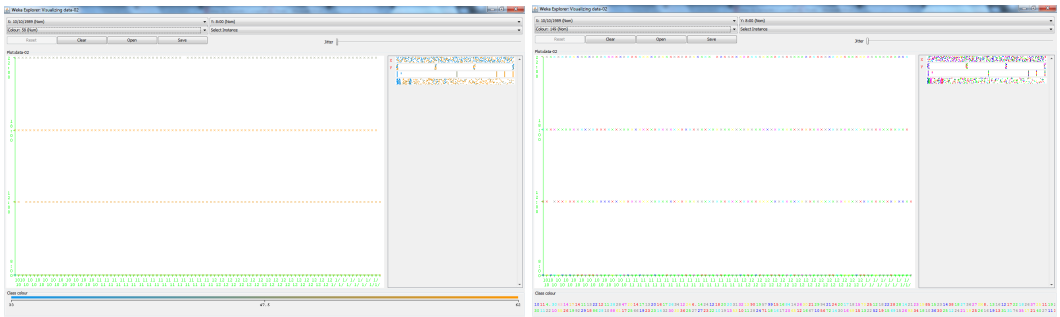
Figure 4: Classification definition based on Code attribute of the Input data



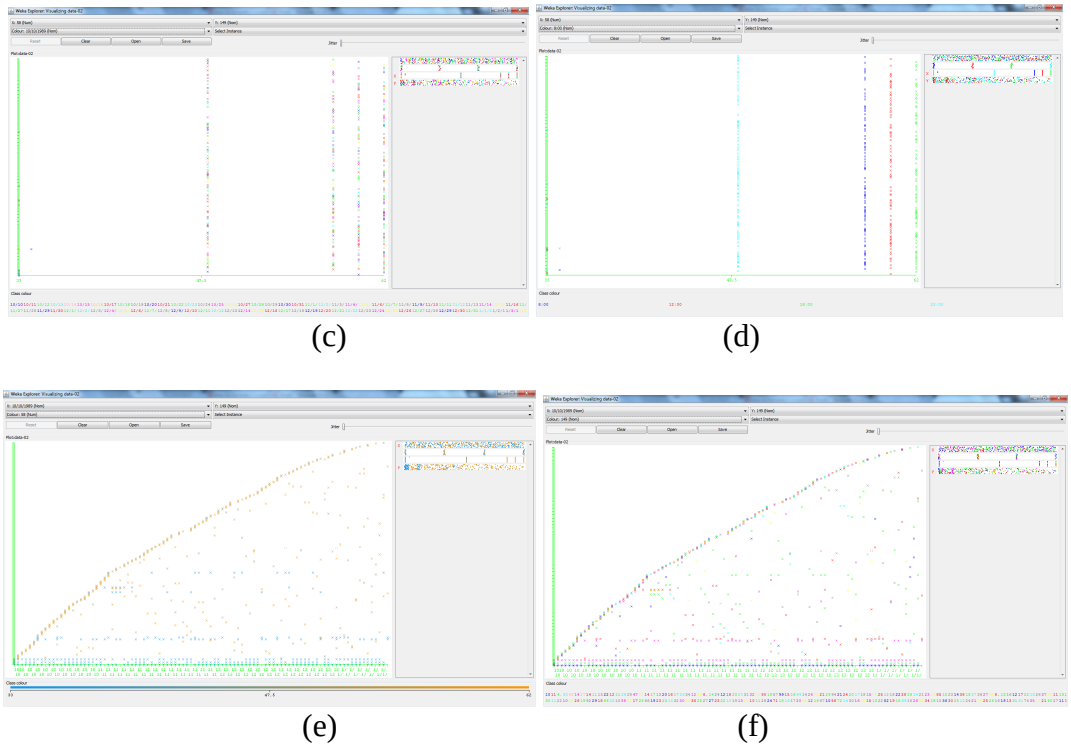
(a) (b)

Figure 5: Classification definition based on Value attribute of the Input data

The overall Graphical representation after the SVM classification with all the dataset attributes on dependency is shown in figure 6 below:



(a) (b)



(c) (d)

(e) (f)

Figure 6: (a) Date, Time and Code, (b) Date, Time and value, (c) Code, Value and Date, (d) Code, Value and Time, (e) Date, Code and Value, (f) Date, Value and Code

Conclusion

The major task of today's world is to detect anomalies which is found to be impossible for due to the current circumstances which seem to be having the most enormous data in almost all the sector one most important sector which has the difficulty to retrieve the anomaly is the medical stream where the number tests carried out and the number of results obtained on these tests are equally likely massive the ranges also seems to be massive, as a result the idea of introduction of anomaly detection algorithm into the field of medical has been stated and has been proven to retrieve resultants considering the diabetics dataset for this experiment was accomplished. The results were obtained using the WEKA tool which is one of the most popular tools amongst the tools of data mining.

References

[1]. H. Jair Escalante, "A Comparison of Outlier Detection Algorithms for Machine Learning", 2005 http://www.researchgate.net/publication/228728521_A_comparison_of_outlier_detection_algorithms_for_machine_learning/links/0912f50b777c20ab5e000000.pdf

[2]. Nyalkalkar, K., Sinha, S., Bailey, M. and Jahanian, F., "A Comparative Study of Two Network-based Anomaly Detection Methods", *INFOCOM, 2011 Proceedings IEEE,* 176 – 180, 2011

[3]. Kevin S. Killourhy and Roy A. Maxion, "Comparing Anomaly-Detection Algorithms for Keystroke Dynamics", *Dependable Systems & Networks, 2009. DSN '09. IEEE/IFIP International Conference,* pp. 125 – 134, 2009

[4]. Mohsen Zare Baghbidi, Kamal Jamshidi, Ahmad Reza Naghsh Nilchi and Saeid Homayouni, "Improvement of Anomaly Detection Algorithms in Hyperspectral Images Using Discrete Wavelet Transform", *Signal & Image Processing : An International Journal (SIPIJ),* Vol.2, No.4, December 2011

[5]. Tarem Ahmed, Boris Oreshkin and Mark Coates, "Machine learning approaches to network anomaly detection", *USENIX Association Berkeley, CA, USA ©2007,* No. 7, 2007

[6]. Bache, K. & Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. *Irvine, CA: University of California, School of Information and Computer Science*. 2013.

[7]. O Mazhelis, "One-Class Classifiers: A Review and Analysis of Suitability in the Context of Mobile-Masquerader Detection", *Joint Special Issue — Advances in end-user data-mining techniques*, Vo. 36, 2006

[8]. Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander, "LOF: Identifying Density-Based Local Outliers", Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000

[9]. Dr. Sudhir B. Jagtap, Dr. Kodge B. G, "Census Data Mining and Data Analysis using WEKA" *(ICETSTM – 2013) International Conference in Emerging Trends in Science, Technology and Management,* 2013

[10]. Charalampos Mavroforakis, "Data mining with WEKA", Boston University CS105, Fall 2011.

[11]. Gowri, S., Anandha Mala, G.S., "Improving intelligent IR effectiveness in forensic analysis", *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, 2012