MLMN: AN EFFICIENT FRAMEWORK FOR MULTI-LAYER MODEL FOR NOVELTY DETECTION IN DATA MINING

[1]ASHAY URADE, [2]POORVA AGRAWAL
[1]Symbiosis International University, Department of Computer Science, Pune
[2]Symbiosis International University, Department of Computer Science, Pune
E-mail: [1]ashay.urade@sitpune.edu.in , [2]poorva.agrawal@sitpune.edu.in

ABSTRACT

Data Classification is one of the boons to the mining process which eventually eases the process of machine learning. But it is always a doubt about the data that is being used for the classification process. As in Supervised, semi-supervised and unsupervised learning styles the selected data for the classification is always having a chance of missing unknown and new data from the data pool this adversely affects the quality of the mining process. So the process of extraction of unknown and new data from the huge data pool for classification process is very important and it is known as novelty detection. So this research of MLMN provides an ideal way to detect the novelty data that is been using in the further classification process. This research mainly focuses on the multiple layers like 1) Pattern identification, 2) Density Classification, 3) Distance identification and 4) Learning models for the efficient process of novelty detection.

Keywords: *K-Means, Gaussian Distribution, Euclidean Distance, Hamiltonian Jacobi, Fuzzy Logic.*

1. INTRODUCTION

This Specification provides a complete description of all the functions and constraints of the “An efficient framework for Novelty detection using level set based approach”. The document describes the issues related to the system and what actions are to be performed by the development team in order to come up with a better solution.

The basic idea of identifying novelty data comes from the fact that while performing classification most of the new or unknown data is been missed from the process. So it affects the quality of the classification. Therefore, the proposed system put forwards an idea of identifying novelty data from huge data sets using Gaussian distribution and fuzzy logic classifier and K-means clustering which are powered with Euclidean distance factor.

Data mining is an essential topic in research field in which it explore knowledge from huge consecutively generated information produced from non-stationary distribution [1]. Novelty data is having the capability to pick out the newest or previously unknown situation. It particularly helps while dealing and analyzing the numerical data in large quantity wherein the object may additionally seem, disappear or evolve over the year. There are many who are doing research on the utility of novelty detection. Here [1] describes new estimation technique for multiclass novelty detection in which information streams consider three aspects which are in following:

1) First of all deal with unsupervised learning which generates novelty pattern without any alliance but providing the proper instructions.

2) Confusion matrix that grows over time.

3) Confusion matrix in which column representing unknown information.

In [2] elaborates the Autonomous system for inspection, protection seeks, guard, and saving someone from the difficult situation are the important major topic of extensive studies. Attentiveness in security and military safety related industry, traffic management, and other security-related institutions. Various kinds of sensors are used in order to detect the certain object to deal such situation. The optical cameras are an important tool because it is more reliable and more effective. Detecting any object is based on the information of color sensitivity. Using this color sensitivity data which is more robust than movement’s indicator of the object. Technique mentioned in the paper [2] is by using a combination of 3D laser and infrared cameras which give the output of grey scale image. In this current situation, the system is available which are having high computational power and high memory storage, but it required human intervention which is not treated as autonomous.

To eliminate the human intervention and to make it autonomous, it requires two important aspects that are novelty detection and monitoring object in streaming of any video. These to aspect are rooted in the recursive strategy which helps to provide computational efficiency and also suitable for real-time scenarios. Original strategies for recursive density estimation (RDE) using Cauchy type kernel is suggest for visible novelty detection and take help of evolving Takagi-Sugeno (ETS) neuro-fuzzy system for monitoring the object observed by the RDE strategies. This suggested strategy is more advanced version for estimation and even more helpful for background elimination particularly in novelty detection.

Bodesheim, Paul, et al elaborate [3] the object detecting samples from past history especially of unknown classes is a vital assignment. And specifically, when real-world applications are the main concerned. [3] Suggest how to use a null-space approach for novelty detection. The lonely model uses more than one known classes and detects novelties for the set of instructions with a single version. This technique utilizes a projection in a common subspace and detects novelty data. In which of all known class which consists training samples has zero intra-class variance. This subspace is also known as NULL space for the training data. This strategy heavily depends on the distance measure but not rely on density estimation. This approach produces an easy and more effective technique for multiclass novelty detection. At the end, we can say that this NULL space strategy is most suitable for multiclass novelty detection.

Bodesheim, Paul, et al. [4] investigated gaining of locality knowledge in multiclass novelty detection. Estimating the newness of a new sample is a very hardest task due to the huge variability of known object groups. The outcome advocates that it's importantly more sufficient to compute novelty ranking with respect to borders of well-known class. In addition, it specified this framework is suited for face detection especially those face which are not known. In huge scale novelty detection, due to the fact that we are also involved kernel-based approach for classes. By selecting proper size of area its computational time will be also optimal.

In [5] Provides Novelty detection is of especially used in the analysis of high righteousness system. To performing novelty detection the support vector machine are a most popular tool and it is a well-known traditional technique to use a train validate test approach and also including cross-validation to choose suitable values for its parameter. The probabilistic technique helpful in traditional method where it is provides facility of selecting a probabilistic novelty calibration approach for one-class SVMs. By this probabilistic method is help to provide some important aspects which are in following:

1) Online novelty detection.

2) Its used in tracking the highly integrity industrial combustion plant.

3) Help to observing the decline in patient physiological condition during patient vital-sign monitoring.

Novelty detection is the approach to find out those data which has not captured by the classifier of data. Mainly classifier is created by applying some rule based approach and this rule based approach only provide those data which do fall under the certain rules. These rules are built according to requirements of the system. So in order to capture those data which does not fall under these rules and those data which are important in such scenario novelty detection is helpful approach.

It takes data set as input first step is to make a Level set function which helps to classify the data .after that second step is boundary evolution try to catch those data which has not captured by the classifier and the last step is boundary termination .

2. RELATED WORK

The approach of conditional multivariate complex Gaussian distribution for distribution state estimation (DSE) [6] offers to calculate the standard derivation and mean of state variables by not having any repetition of the process. This technique involves certain aspects such as charge correlation, load that is uncertain and measuring delusion. The conditional multivariate complex Gaussian distribution for DSE is practically experimented on bus dispensation web. In this technique first of all it provides the bus voltage, branch current through multivariate complex Gaussian distribution (MCGD) and also help to provide the direct load flow and a linear transformation. And after the experiment, it specified in a simulation that increases the effectiveness as compared with regular weighted least square (WLS) based DSE. This technique is especially suitable for smart network distribution which consist the real time DSE.

[7] Narrates the technique Gaussian pseudo-random number generator which is capable of coming close to a Gaussian distribution with quality which is up to the mark. And the most importantly it provides us optimal time complexity.

Mezache, Amar, et al.[8] suggested the technique of compound inverse Gaussian combination of Rayleigh pdf and IG pdf and it has three criterion and In additional the thermal sound is added in order to get the proper model to depict the real sea clutter statistic .It make distribution in such way which is more workable .All experiments conclusions suggest that compound inverse Gaussian model is very well suitable for sea clutter with respect to the real data .This technique is prime suitable for real data, it helps us to put appropriate observation gateway. And also helps to provide an appropriate complementary cumulative distributed function.

In this [9] method, it consists of a top rated detector for the multiplicative watermarking plan in contourlet domain with the help of normal inverse Gaussian distribution. This proposed methodology is consisting the Bayesian log likelihood ratio criterion through the valuable restriction theorem, the PDF of the log likelihood ratio has been considered as Gaussian and after performing several experiments on this top rated detector and also get compared with several other detectors. The result of all of these experiments and testing concludes that the normal inverse Gaussian based watermark detector has got advanced detection quality as compared to other detectors particularly in the problem of blind watermark detection in contourlet domain . When we measured all quality of all detectors this normal inverse Gaussian based watermark detector is more durable.

As per as clustering is concerned K-means algorithm is well known for its adaptability and easiness [10]. But it has a certain limitation that is the initially value of K should be specified .This paper [10] include various available techniques about how to initialize the value of 'K', this value of K equal to a number of cluster. The new technique has been taken by assuming more than one value of K for users to deal different scenario in different stages .And it also helps to give different results from different stages. Most importantly it will be computationally economical if it does not have a large data set.

In order to make clusters after classification from huge data set the distance measure is a factor which plays a huge role and there are different kind of distance measure available to deal certain required condition while gathering data to make clusters, for example, Euclidean distance and Manhattan distance. But here we are using Euclidean distance because it is easiest distance measure technique and in this scenario we have to initial K value which less than 64. Importantly it helps to have less time complexity and if the condition is where the value of K is greater than 64 then Manhattan distance measure is much better option. The time complexity is calculated by finding a number of iteration perform in that function. Which is shown in following table:

Table 1: Time complexity analysis on basis of number of loops executed in function

K values	Number of time loop execute in Euclidean distance	Number of time loop execute in manhattans distance
8	4	8
16	5	6
32	4	4
64	4	3

Hamerly, Greg, and Charles Elkan. [11] Treated the clustering in high dimensions is an up most problem in past few years. As per as research is concerned in this area researchers had used dimensionality reduction approach and then forms the clusters and after these two steps the algorithm is used which is comes under low dimensional clustering.

[12] This paper explain about how quickly this approach is solving a Hamiltonian-Jacobi equations with Dirichlet boundary conditions which come under a class of time independent. This approach is a combination of tracing property with Godunov construction and

gauss seidel iterations. By the easiest structure of the convexity they specify a simply reduced expression for the Godunov Hamiltonian. With this reduced Godunov flux ,the complication of calculation is come down to only eight cases in two dimension space which is specified in the expression into easy reduced Gauss seidel type iteration procedure .Here the observation suggests the time complexity for convergence is O(n) in which n is a total number of grid points.

[13] Says Hamiltonian Monte Carlo (HMC) is a Markov chain Monte Carlo (MCMC) algorithm that not follows the random walk pattern. Hamiltonian Monte Carlo effectiveness is highly dependent on the two parameters first is step size and the second parameter is a number of steps which is denoted by S. Suppose if value of ‘S’ is very small then output produced by the algorithm will be not desired random walk pattern and if value of ‘S’ is very big then algorithm will wastes more time in computation. So to deal this problem in this paper [9] they explained about No-U-Turn sampler (NUTS).No-U-Turn Sampler is a modified and upgraded version of Hamiltonian Monte Carlo where it is not necessary to give a number of step ‘S’ NUTS provide more effectiveness and also provide the human interaction. The main thing in NUTS is it considers the step size on the fly based on primal-dual averaging.

[14] It include the fuzzy logic is helpful to the formalization of two important ability of human behavior. Fuzzy logic plays an important role where the decision making is vital in a complex scenario .A complex scenario means where the data is uncertain, less available information, scenarios of the dilemma.

By using almost negligible computation process it can perform the physical task and also decision-making task. Fuzzy logic is upgraded form of an intellectual system. Relational facet is played a role in almost all practical application of fuzzy logic.

The text summarization are divided into two methods and these are extraction and abstraction [15]. The main objective of text summarization is to extract the highlighted main sentence of paragraph or document with the help of extracting method. They extracted the main properties of sentence length, term consists title feature, sentence length, term weight, sentence similarity, proper noun.

By using a level set method which is having a small boundary and this boundary are static. Mainly Hamiltonian Jacobean Skelton used to find the shocks particularly in such system where the shock is changing constantly. Still this method conserves the shock .But finding the actual shock is not fully clearly expressed. In another side Hamiltonian method is basic building block for Hamiltonian physics.Which is vitally important aspect of shock theory.

3. PROPOSED SYSTEM

In this section, we describe our framework for Novelty detection method using a level set function with the below-mentioned steps as shown in figure 1.

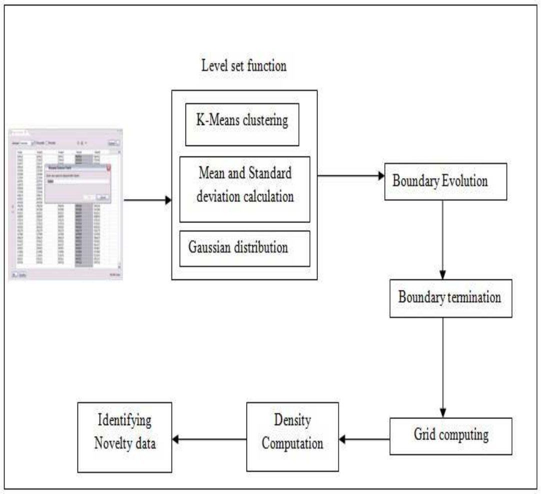

Figure 1: Overview of our approach

Step 1: Very first step of the system is to select the dataset which is in text format. This dataset contains total 9 attributes. Out of these 9 attributes, only four important attributes are selected for the further implementation. For each of the row of the dataset, a Euclidean distance is calculated.

Once the Euclidean distance is calculated all the rows are sorted according to the Euclidean distance. Smallest and the largest value are extracted from this sorted Euclidean distances. This extracted smallest and largest value is then divided into five parts. Then row will be moved to the cluster by checking the occurrence of distance in a particular range. In this way, clusters are created.

Step 2: Once the clusters are created mean and standard deviation of each of the four selected attribute is calculated. So for each cluster, we have four mean and four standard deviations. So from this mean and standard deviation, the Gaussian value is calculated for each of the attributes.

Step 3: So depending on the mean, standard deviation and Gaussian minimum and maximum range of each attribute is find out. And then for each of the row, its occurrence is checked for each minimum and maximum range. In this way, each cluster is further divided into four clusters. These newly generated clusters are known as Neurons.

(A) This process is done using Gaussian distribution factor or normal distribution function Gaussian Distribution Equation

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \quad \ldots\ldots\ldots\ldots (1)$$

Where,
μ= mean of distribution
$\sigma 2$ = variance of distribution
x= continuous variable
g(x) = probability of x

Step 4: Boundary termination - Once neurons are generated we have to find out the cluster boundary. So to accomplish the task minimum and maximum value of each of the neuron for all the four attributes are found out. This minimum and maximum value denote the boundary of that particular cluster.

5: Once the Boundary termination is over then by calculating the density of the data within the newly revised boundaries classification rules are been estimated. Then data density computation is calculated to identify the fine novelty data form the dataset.

The whole process can be depicted in the below algorithm:

ALGORITHM 1: NOVELTY DETECTION

```
Input: Dataset Containing Set of attributes
Output: Cluster Boundary
Step 1: Start
Step 2: Fetch Diabetes dataset $D_s$
Step 3: Find 5 clusters C1, C2, C3, C4, C5 from $D_s$ using K means
Step 4: Find sd, mean and Gaussian function of each cluster by considering four more important attributes i.e. d1,d2,d3,d4
Step 5: find minimum range and maximum range of each clusters for four attributes
Step 6: minimum range =mean
If (Gaussian value > (mean * 2))
        Maximum range=mean + sd
Else
        Maximum range =mean + Gaussian function
Step 7: Apply ANN on C1, C2, C3, C4, C5 to generate more clusters by using minimum range and maximum range
Step 8: store all the newly generated clusters to $N_C$
Step 9: Set Fuzzy parameters
Step 10: For i=0 to N (where N is length of Nc)
Step 11: for each $Nc_i$ find minimum and maximum value of clusters as a boundary of that cluster
If (more than 2 attributes are matched).
Step 12: End for.
Step 13: End for.
Step 14: Novelty data detection using density calculation
Step 13: Stop
```

4. RESULTS AND DISCUSSIONS

Some experimental evaluations are performed to show the effectiveness of the system. And these experiments are conducted on windows based java machine with universally used IDE NetBeans. Also the numbers of retrieved novelty data form the data method set is used to set benchmark for performance evaluation.

Numbers of relevant retrieved novelty data from the dataset is used to show the effectiveness of the system

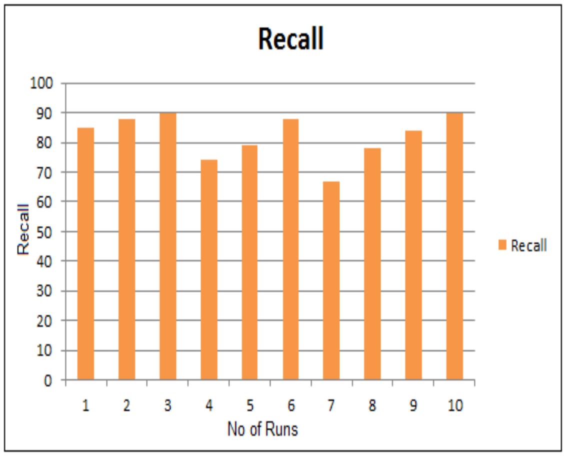

Figure.2. Average precision of the proposed novelty detection

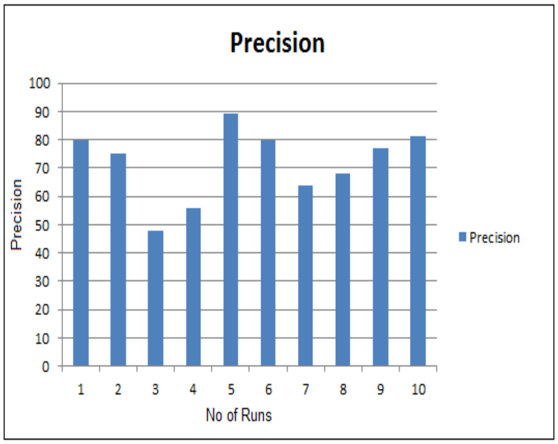

Figure.3. Average Recall of the proposed novelty detection method

Below are the definition of the used measuring techniques i.e. precision and recall.

Precision: It is a ratio of numbers of proper novelty data retrieved to the sum of total numbers of relevant and irrelevant novelty data retrieved. Relative effectiveness of the system is well expressed by using precision parameters

Recall: It is a ratio of total numbers of relevant novelty data retrieved to the total numbers of relevant documents not retrieved. Absolute accuracy of the system is well narrated by using recall parameter

Numbers of scenarios presents where one measuring parameter dominates the other. By taking such parameters into consideration we used two measuring parameters such as precision and recall.

For more clarity let we assign

- X = the number of relevant novelty data retrieved,
- Y = the number of relevant novelty data retrieved are not retrieved, and
- Z = the number of irrelevant novelty data are retrieved.

So, Precision = (X/ (X+ Z))*100

And Recall = (X/ (X+ Y))*100

In Fig. 2, by observing figure 2 it is clear that the average precision obtained by using novelty detection method is approximately 71.8%.

The Fig. 3, shows that the system gives 82.3% recall for the novelty detection method. By comparing these two graphs we can conclude that the novelty detection method gives high recall value compare to the precision value.

5. CONCLUSION

In our proposed approach of novelty detection, we used the dataset belongs to diabetic's patients. Here in the processing system efficiently creates the level sets by considering the min and max values which are defined using the probabilistic approach of the normal distribution. Then by revising new boundaries, an enriched classification protocols are been creating which eventually helps to identify the best novelty data.

REFRENCES

[1] Ribeiro de Faria, Elaine, et al. "Evaluation of Multiclass Novelty Detection Algorithms for Data Streams." *Knowledge and Data Engineering, IEEE Transactions on* 27.11 (2015): 2961-2973.

[2] Angelov, Plamen, Ramin Ramezani, and Xiaowei Zhou. "Autonomous novelty detection and object tracking in video streams using evolving clustering and Takagi-Sugeno type neuro-fuzzy system." *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*. IEEE, 2008.

[3] Bodesheim, Paul, et al. "Kernel null space methods for novelty detection."*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013

[4] Bodesheim, Paul, et al. "Local novelty detection in multi-class recognition

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

problems." *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 2015.

[5] Clifton, Lei, et al. "Probabilistic novelty detection with support vector machines." *Reliability, IEEE Transactions on* 63.2 (2014): 455-467

[6] Arefi, Ali, Gerard Ledwich, and Behnaz Behi. "An Efficient DSE Using Conditional Multivariate Complex Gaussian Distribution." *Smart Grid, IEEE Transactions on* 6.4 (2015): 2147-2156.

[7] Condo, C., and W. J. Gross. "Pseudo-random Gaussian distribution through optimised LFSR permutations." *Electronics Letters* 51.25 (2015): 2098-2100.

[8] Mezache, Amar, et al. "Model for non-rayleigh clutter amplitudes using compound inverse gaussian distribution: an experimental analysis."*Aerospace and Electronic Systems, IEEE Transactions on* 51.1 (2015): 142-153.

[9] Sadreazami, H., M. Omair Ahmad, and M. N. S. Swamy. "Optimum multiplicative watermark detector in contourlet domain using the normal inverse Gaussian distribution." *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on*. IEEE, 2015.

[10] Pham, Duc Truong, Stefan S. Dimov, and C. D. Nguyen. "Selection of K in K-means clustering." *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 219.1 (2005): 103-119.

[11] Hamerly, Greg, and Charles Elkan. "Learning the k in A> means." *Advances in neural information processing systems* 16 (2004): 281..

[12] Siddiqi, Kaleem, et al. "Hamilton-jacobi skeletons." *International Journal of Computer Vision* 48.3 (2002): 215-231.

[13] Homan, Matthew D., and Andrew Gelman. "The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo." *The Journal of Machine Learning Research* 15.1 (2014): 1593-1623.

[14] Zadeh, Lotfi A. "Is there a need for fuzzy logic?." *Information sciences*178.13 (2008): 2751-2779.

[15] Suanmali, Ladda, Naomie Salim, and Mohammed Salem Binwahlan. "Feature-Based Sentence Extraction Using Fuzzy Inference rules." *2009 International Conference on Signal Processing Systems*. IEEE, 2009.

[16] Clifton, David A., et al. "Bayesian extreme value statistics for novelty detection in gas-turbine engines." *Aerospace Conference, 2008 IEEE*. IEEE, 2008.

[17] Willett, R. M., and Robert D. Nowak. "Minimax optimal level-set estimation."*Image Processing, IEEE Transactions on* 16.12 (2007): 2965-2979.

[18] [19] Tuceryan, Mihran, and Anil K. Jain. "Texture analysis." *Handbook of pattern recognition and computer vision* 2 (1993): 207-248.

[19] Gogoi, Prasanta, et al. "A survey of outlier detection methods in network anomaly identification." *The Computer Journal* (2011): bxr026.

[20] Gupta, Madhu, et al. "Top-k interesting subgraph discovery in information networks." *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*. IEEE, 2014.

[21] Görnitz, Nico, et al. "Toward supervised anomaly detection." *Journal of Artificial Intelligence Research* (2013).

[22] Song, Le, Choon H. Teo, and Alex J. Smola. "Relative novelty detection."*International Conference on Artificial Intelligence and Statistics*. 2009.

[23] Blanchard, Gilles, Gyemin Lee, and Clayton Scott. "Semi-supervised novelty detection." *The Journal of Machine Learning Research* 11 (2010): 2973-3009.