The International Journal of Engineering and Science (IJES)
|| Volume || 6 || Issue || 5 || Pages || PP 77-81 || 2017 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Anomaly Detection in Temporal data Using Kmeans Clustering with C5.0

Mani Mehrotra[1], Nakul Joshi[2]
[1]Department of Computer Science, DIT University Dehradun
[2]Department of Electronic Communication, DIT University Dehradun

ABSTRACT

Anomaly detection is a challenging problem in Temporal data .In this paper we have proposed an algorithm using two different machine learning techniques Kmeans clustering and C5.0 decision tree , where Euclidean distance is used to find the closest cluster for the data set and then decision tree is built for each cluster using C5.0 decision tree technique and the rules of decision tree is used to classify each anomalous and normal instances in the dataset .The proposed algorithm gives impressive classification accuracy in the experimented result and describe the proposed system of kmeans and C5.0 decision tree

Keywords*: Anomaly detection, K-Means clustering, C5.0 decision tree, Information gain, Decision tree*

Date of Submission: 19 May 2017 Date of Accepted: 31 May 2017

I. INTRODUCTION

Anomaly detection is the identification of patterns which are abnormal in nature and does not conform to expected pattern or event [1]. An anomaly is something that deviates from what is standared normal or expected. Anomaly detection is important for several application domains such as public, health, climate studies. Defining a normal region is a challenging problem as the exact notation of an anomaly is different for different application domain[7] .Anomalies are divided into three parts Point anomalies, Contextual anomalies ,Collective anomalies and these types of anomalies are commonly in Spatial, sequential or temporal data[7]. Many anomaly detection techniques have been specifically developed for certain application domain. In this paper we have use temporal data set, it is a temperature related data which contain different temperature values for an area, and in this data set we try to find out Contextual anomalies. And for this we are using Kmeans with C5.0decision tree algorithm. In this approach Kmeans clustering is used initially to partition dataset into K closest cluster. And after that apply C5.0 technique to built decision tree for each closest cluster and the rules created by decision tree are used to detect anomalies in the dataset.

The following section of the paper are organized as follows section (2) will describe related work on the field of anomaly detection, Kmeans, C5.0 decision tree section(3) will describe the proposed algorithm section(4) will describe the Experimental result and finally Section(5) will describe conclusion and future work

II. RELATED WORK

Extensive research has been done in the field of anomaly detection, various techniques are there for detecting anomalies in a dataset. In this paper two types of machine learning algorithms are used kmeans and decision tree C5.0 which is use for detecting unexpected patterns in a dataset.[22]compared different clustering algorithms which conclude that clustering is an efficient approach for anomaly detection[23]simple Kmeans is time efficient and data set using Kmean perform better result among all the four clustering algorithms but there are some limitation in Kmeans as it work only for well shaped cluster and fixed number of cluster can make it difficult to predict the value of K ,which can be overcome by modifying Kmeans algorithm,[24]modified kmeans overcome the problem of finding the optimal number of clusters and drawback of this approach is ,it's time complexity is more than K means for larger data sets [6] kmeans has been combined with various other techniques such as apriori algorithm,ID3, decision tree for better accuracy, Kmeans with apriori is more resist to noise or outliers.[16] proposed an technique by cascading Kmeans with different classification techniques, this removes the anomalies from Kmeans using id3, it overcome the disadvantage of both ID3 and Kmean but integrating Kmeans +id3 is a time consuming process.[8]:proposed an anomaly detection method by combining Kmeans clustering technique and C4.5 decision tree method ,this method achieve better performance in comparison to Kmean , C4.5,KNN .[11] Help us in understanding sequence anomaly detection problem and how existing Kernel based Window based,Markovian,Hidden Monrovian Model based techniques related to each other . What are their strength and weakness, how effectively techniques solve a problem for which they are not initially intended.[13] has comparatively evaluated various techniques and conclude that performance depend on the nature of the sequence and nature of anomaly in the sequence data set ,no technique can be label as best technique. [26] proposed a hybrid approach for anomaly detection in large scale dataset using meta heuristic method .this approach show a better accuracy in generating a suitable number of detector when compared to algorithms like Navies Bayis, Decision tree, Multilayer feedback neural network, Bayes Network ,Bayesian Logistic Regression, Radial Basis function Network. [4] compared ID3, C4.5, and C5.0 decision tree classifiers with each other and conclude that C5.0 give more accurate and efficient result among all classifiers. [25] propose a new form of SVM approach for the multiclass classification and anomaly detection in a single step which improves the performance if the data contain vectors which is not represented in the training dataset but the problem is the difficulty in setting one of the algorithm's parameters.[2]uses C5.0 and one class SVM for building an anomaly detection technique. In this paper we have used Kmean and C5.0 the reason for choosing Kmean is time complexity O(nkm) where n is the number of clusters, k is the number of patterns m is the no of iterations , its space complexity O(n+K) and its scalibity ,its order independent. The reason of choosing C5.0 is it is more efficient, its decision tree is smaller in cooperation with C4.5, unnessary attributes have been automatically removed by C5.0

2.1 Kmean algorithm

It is a partitional clustering approach which is used to cluster numerical data. The partitioning clustering method partitions the element into the set of non-overlapping and un-nested or one level clusters, so as to maximize the evaluation value of clustering where each cluster optimizes the clustering criterion. K means clustering is used to cluster the group of objects into k disjoint groups based on their attributes .The objects within the cluster will be the similar while those from different cluster differ from one another .In k means ,we define the two measures ,distance between the two points and the distance between the two clusters

Euclidian distance is the most popular method to measure the distance. the formula for the euclidian distance is :

p=$(p_{1,}p_{2,}p_{3})$ and Q=$(q_{1,}q_{2,}q_{3})$ are the two points in Euclidian space

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+(p_3-q_3)^2}$$

Input: numerical, There must be a distance metric defined over the variable space
->euclidian distance
Output: the centres of each discovered cluster, and the assignment of each input datum to a cluster.
-> centroid

The algorithm for the k means clustering is :
step 1: Choose K, then select K random "centroids "
step 2: assign records to the cluster with the closest centroid
step 3: Recalculating the resulting centroids
step 4: repeat step 2 and 3 until record assignments no longer change.

The reason to choose K means clustering is that the k means clustering can handle large datasets. Additionally, observations are not permanently commited to a cluster .They are moved when doing so improves the overall solution

2.2 C 5.0 decision tree technique:

The C 5.0 is the machine learning algorithm developed by Quinlan based on decision tree[2]It is an extension of C4.5, It is better than the C4.5 on the efficiency and the memory .In C5.0 model samples are split on the basis of biggest information gain field[3].C5.0 generates classifiers expressed as decision trees, but it can also construct classifiers in more comprehensible ruleset form.

Improvement in C5.0 from C4.5 algorithm:

C5.0 decision tree is smaller in comparison with C4.5. In unseen cases the C5.0 rule set have lower error rates [4].It is easier to understand C5.0 model in comparison to some other type of model and it required lesser training time to estimate [2]. The scalability of both decision tree and rule set is greatly improved . Advantages of computers with CPUs and/or cores can also be taken by C5.0 [6]

2.2.1 Decision tree:

decision tree are a flexible method very commonly deployed in data mining applications . In this paper we will focus on the decision tree used for classification problems.

There are two types of trees classification and regression tree.

Classification trees: observations are used to segment into more similar groups. They usually apply to outcomes that are categorical or binary in nature
Regression trees: are variations of regression and each node returns its average value at each node. Regression trees can be applied to outcomes that are continous [17]

2.2.2 Information Gain:
The training data is separated by using an well define attributes. It is based on the entropy measure commonly used in information theory [2].It is defined as the differnce between the base entropy and the conditional entropy of the attribute.
So the most information attribute with highest information gain [2]
Let T is the training dataset
X(c) is the class I where c=1,2,3…n

$$I\left(T_1,T_2...T_n\right) = -\sum p_c \log_2 p_c$$

T_c is the number of samples in c

$$p_c = \frac{T_c}{T}$$

$\log_2$ is binary logarithm let attribute A has v distant values
Entropy=E(A) is

$$\sum \{ \frac{(T_{1j} + T_{2j} + ... + T_{nj})}{T} \} * I(T_1, T_2, ...T_n) \qquad \text{j=1}$$

where Tcj is the sample in class c and subset j of attribute

$$I(T_{1j}, T_{2j}, ...T_{nj}) = -\sum p_{cj} \log_2 p_{cj}$$

$$gain(A) = I(T_1, T_2, ...T_n) - E(A) \qquad \text{Eq(1)}$$

The algorithm for C5.0 decision tree is [17]:
step 1: The C5.0 generates a either a decision tree or a ruleset
step 2:Pick the most informative attribute
step 3: Find the partition with the highest infomation gain using Eq (1)
step 4: at each resulting node ,repeat step1 and 2

III. PROPOSED ALGORITHM

We proposed an anomaly detection algorithm by using two machine learning algorithm Kmeans and C5.0 .initally Kmeans is used for partitioning the dataset into K closest cluster using Euclidean distance formula and than C5.0 techniques is applied on each closest cluster to built decision tree for each cluster and classify the each instance into normal or anomaly using decision tree result
The algorithm consist of two phases selection phase and classification phase

1) Selection phase: the closest cluster is selected for each test instance. In the selection cluster the decision tree corresponding to the cluster is generated
2) classification phase :The test instance is classified into normal and anomaly using the C5.0 decision tree result and the cluster label as normal or anomaly

Algorithm
Test instances $z(i)$ where $i = 1 \dots n$
1). Read the data set
2). Select K initial centroid of the cluster randomly
3).for each instance $z(i)$in the data set, find the closest cluster using eucledin distance $d(z(i), c(j)) \quad , j = 1....k$

$$d(z(i), c(j)) = \sqrt{\sum_{a=1}^{m} (z_{ia} - c_{ja})^2}$$

4) Converting the variables into factorial and integer values for the better result
5) Compute C5.0 algorithm for the closest cluster using highest information gain

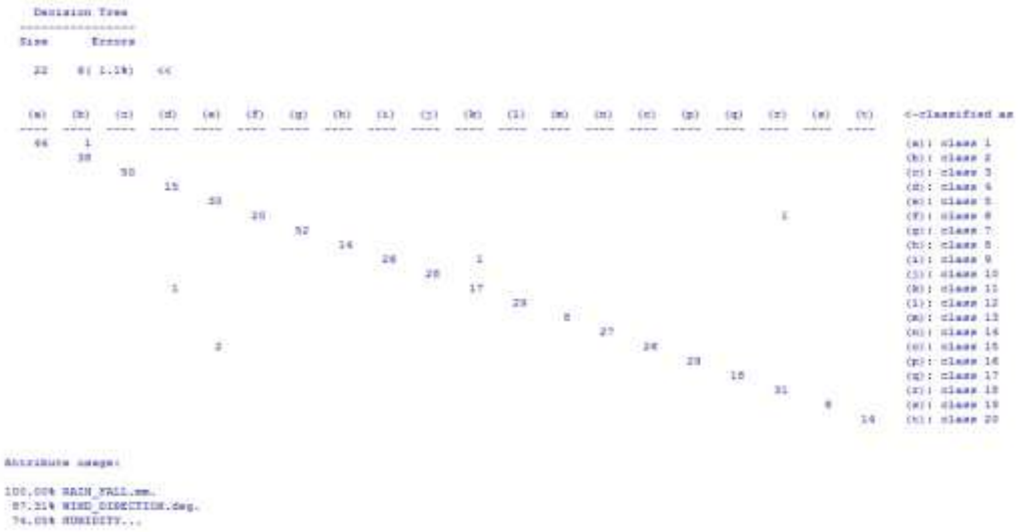
Using Eq $gain(A) = I(T_1, T_2, ...T_n) - \sum \{\frac{(T_{1j} + T_{2j} + T_{nj})}{T}\} * I(T_1, T_2, ...T_n)$ j=1

6). Apply the test instance $z(i)$ over C5.0 decision tree of the computed closest cluster
7).classify the test instance $z(i)$ as normal 0r anomaly using the decision tree and include it into the cluster
8).update the cluster center
9). End

From step 2to5 is selection phase and from step 6to8 is classification Phase

IV. EXPERIMENTAL RESULT

This section will present the performance of proposed algorithm and for this we have used ISRO dataset. the dataset have 16 attributes but we have taken only 7 attributes for classification of instance into normal and abnormal, we apply Kmeans on the dataset for portioning the dataset into K clusters , here we have taken K=20 ,number of iteration =10 if there is overlapping anomaly type of data cannot eliminated by kmeans . so C5.0 technique is use to classify each cluster we have made a new data frame name as final which contain 8 variables from which we have taken clustering variable as target value and partition the dataset into two parts train and test set ,in this paper we take the 90% values for the train set and 10%value for the test set compute the C5.0 for the closest cluster and the result of Kmeans and C5.0 classify each instance into normal or anomaly data



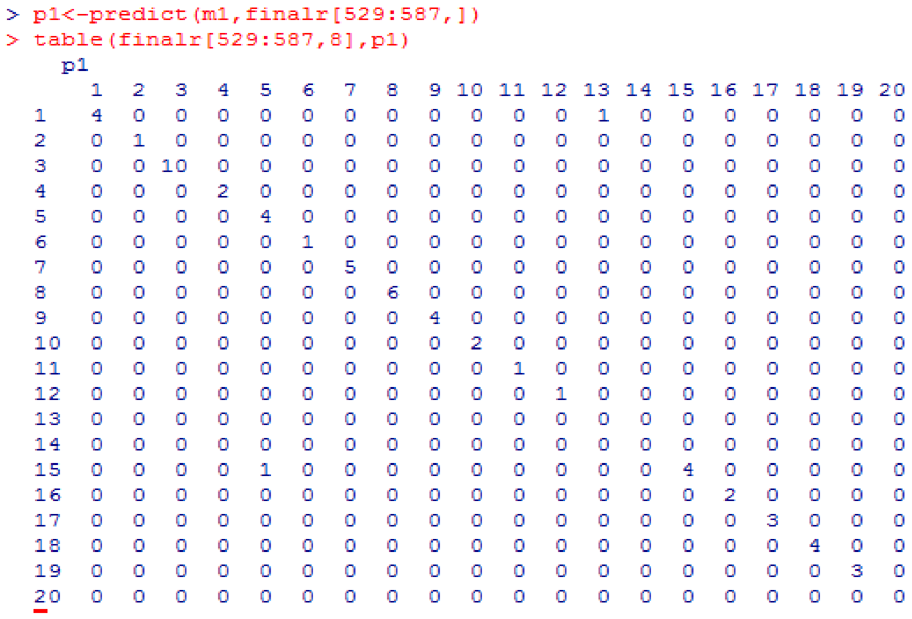
```
 Decision Tree
 ----------------
 Size      Errors

   22    8( 1.1%)   <<

  (a)  (b)  (c)  (d)  (e)  (f)  (g)  (h)  (i)  (j)  (k)  (l)  (m)  (n)  (o)  (p)  (q)  (r)  (s)  (t)    <-classified as
 ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ----
   46    1                                                                                             (a): class 1
        38                                                                                             (b): class 2
             50                                                                                        (c): class 3
                  15                                                                                   (d): class 4
                       43                                                                              (e): class 5
                            20                                                        1                (f): class 6
                                 52                                                                    (g): class 7
                                      14                                                               (h): class 8
                                           26         1                                                (i): class 9
                                                20                                                     (j): class 10
                   1                                 17                                                (k): class 11
                                                          23                                           (l): class 12
                                                               8                                       (m): class 13
                                                                   27                                  (n): class 14
                        2                                               24                             (o): class 15
                                                                             23                        (p): class 16
                                                                                  18                   (q): class 17
                                                                                       31              (r): class 18
                                                                                             8         (s): class 19
                                                                                                 14    (t): class 20


Attribute usage:

100.00% RAIN_FALL.mm.
 87.31% WIND_DIRECTION.deg.
 74.05% HUMIDITY...
```



```
> p1<-predict(m1,finalr[529:587,])
> table(finalr[529:587,8],p1)
    p1
      1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
  1   4  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0
  2   0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  3   0  0 10  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  4   0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  5   0  0  0  0  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  6   0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  7   0  0  0  0  0  0  5  0  0  0  0  0  0  0  0  0  0  0  0  0
  8   0  0  0  0  0  0  0  6  0  0  0  0  0  0  0  0  0  0  0  0
  9   0  0  0  0  0  0  0  0  4  0  0  0  0  0  0  0  0  0  0  0
  10  0  0  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0
  11  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0
  12  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
  13  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  14  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  15  0  0  0  0  1  0  0  0  0  0  0  0  0  0  4  0  0  0  0  0
  16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  0  0  0  0
  17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  3  0  0  0
  18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0  0
  19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  3  0
  20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

Figure2 and 3 shows the classification matrix and confusion matix after applying Kmeans and C5.0 on the training dataset and test data it shows that in training data there are 6 incorrect classifications and 522 correct classification and in test data there are 57 correct classification and 2 incorrect classification the values which are diagonally present in the matrix are showing us the correct pattern and the values which are present on the upper triangle and lower triangle of the matrix are giving us unexpected pattern

V. CONCLUSION AND FUTURE WORK

In this paper we have proposed an algorithm using two different machine learning techniques Kmeans and C5.0 decision tree techniques for detecting anomalies in temperature related dataset the Kmeans is first applied to partition the dataset into K clusters and then C5.0 decision tree is built on each cluster for better classification of instances ,the C5.O decision tree and cluster labels are used to classify the instances as normal and anomaly .our future work is to improve the accuracy by combining different clustering algorithms such as Hierarchical clustering, adaptive resonance(ART)neural network and kohonen's self _organizing maps. With decision tree C5.0. While applying the above algorithm we have seen that our training dataset show no error but in our test data set 393 variables which are correctly predicted and19 variables are incorrectly predicted.

REFERENCES

[1] Michael A. Hayes, Miriam A.M. Capretz "Contextual Anomaly Detection in Big Sensor Data", IEEE Big data 2014 june 27-july 2,2014,Anchorage,Alaska,USA

[2] Meesala Shobha Rani and S.Basil Xaviar" A Hybrid Intrusion Detection System Based on C5.0 Decision tree and one –class SVM" International journal of current Engineering and Technology Vol5.No3(June 2015)

[3] Nilma patil,Rekha Lathi and Vidyachitre "Comparison of C5.0 &CART classification algorithms using pruning technique"(IJERT) vol 1,Issue 4,June 2012

[4] Rutvija pandya,jayati pandya"C5.0 algorithm to improved decision tree with feature selection and Reduced Error pruning " International Journal of computer Application (0975-8887) volume 117-No.16, May 2015

[5] M. Jianliang, S. Haikun, and B. Ling, "The application on intrusion detection based on K-means cluster algorithm," in Proceedings of the International Forum on Information Technology and Applications (IFITA '09), vol. 1, pp. 150–152, IEEE,Chengdu, China, May 2009

[6] XindoWu, Vipin kumar,J.Ross Quinlan ,etal,"Top algorithm in data mining "published online 4 dec2007 springer-verlag London Limited 2007 DOI 10.1007/s10115-007-0114-2

[7] Vchandola, A.Banerjee and V.kumar "Anomaly detection: A survey" ACM comput.surv. vol 41,no 3 pp15:1-15:58 jul 2009

[8] AmuthanprabakarMuniyandi,R.Rajeswari, R.Rajaram"Network Anomaly Detection by Cascading K-Means clustering and C4.5 Decision tree algorithm" International conference on communication and system design 2011.Available online at WWW.science direct.com

[9] Manish Gupta,Jing Gao,etal"Outlier detection forTemporal data: A survey"IEEE Transactions on Knowelge and data Engineering, vol25,no1, January 2014

[10] Hesam Izakian "Anomaly Detection and Characterization in Spatial Time Series Data: A Cluster _Centeric Approach" IEEE Transactions on Fuzzy System (volume22,Issue6)24 January 2014

[11] V.chandola,A.Banerjee, V.kumar"Anomaly detection for discreate sequences: A survey",IEEE Trans,Knowl,Data Eng,Vol 24,no.5,pp.832-839 May 2012

[12] H. Izakian ,W.pedrycz,I. Jamal "Clustering Spatio-temporal data: An argumented Fuzzy-C-Means,IEEE trans. Fuzzy syst ,Vol 21,no5,pp 855-868 oct 2013

[13] V.Chandola,V.Mithal,V.Kumar," Comparavative evaluation of anomaly detection techniques for sequence data" 8th IEEE Int.conf.on Data mining,pisa ,Itly 2008 pp 743-748

[14] Soumi Ghosh, Sanjay Kumar Dubey" Comparative Analysis of Kmeans and Fuzzy C means Algorithm" IJACSA vol 4,no.4,2013

[15] Tejwant Singh, Manish Mahajan**"**Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm ",ijarcsse 2015. Volume 4 ,Issue 5,May 2014 www.ijarcsse.com

[16] K. Hanumantha Rao,etal"Implementation of anomaly detection technique using machine Learning Algorithms" International journal of computer science and telecommunication[volume2, Issue 3,june 2011]

[17] EMC2 Provien Professional copyright 2012 EMC corporation all right reserved

[18] Deepti sisodia,Lokesh singh,etal"clustering techniques:a brief survery of different clustering algorithms"(IJLTET) issue sep 2012

[19] Sweata KJ,Sunita Guruprasad"blocking misbehaving users by Kmean clustering algorithms"(ijircce) vol 3,issue 5may 2015

[20] Ravi Ranjan,G Sahoo"A new clustering approach for anomaly instrusion dection" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.4, No.2, March 2014

[21] Osama Abu Abbas" comparsion between data clustering algorithm "The international Arab journal of Information Technology vol. 5 No.3 july 2008

[22] Sarita Tripathy, etal."A survery of different methods of clustering for anomaly detection" International Journal of Science and Engineering Research Vol.6, Issue 1,Jan 2015

[23] Peerzada Hamid Ahmad"performance evalution of clustering algorithms using different data set", Journal of Information Engineering and Applications

[24] www.iiste.org ISSN 2224-5782 (print) ISSN 2225-0506 (online) Vol.5, No.1, 2015

[25] Ahamed shafeeq ,Hareesha"Dynamic clustering of data with modified Kmeans algorithm" 2012 International Conference on Information and Computer Networks (ICICN 2012) IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore

[26] A. Shilton, S. Rajasegarar, and M. Palaniswami, "Combined multiclass classification and anomaly detection for large-scale wireless sensor networks," in Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on, 2013, pp. 491-496.s

[27] TamerF.Ghanem,etal(2014),"A Hybrid approach for efficient anomaly detection using meta heuristic methods, Journal of Advanced Research.2090 -1232© 2014 Production and hosting by *Elsevier* B.V on half of Cairo University. http://dx.doi.org/10.1016/jare.2014.02.009