AN ENHANCED APPROACH FOR OUTLIER DETECTION AND CLASSIFICATION IN CATEGORICAL DATA USING CLASSIC K-NN ALGORITHM

Monika.M[1], C.P.Balasubramaniam[2] M.Sc(CS).,M.Phil

[1]Research scholar,M.Phil(Computer Science),
Kongu Arts and Science College (Autonomous), Erode-638017.
[2]Assistant Professor, Department of Computer Science,
Kongu Arts and Science College (Autonomous), Erode-638017.

Abstract— In this Paper analysis the healthcare dataset for classifying the outliers into different categories using K-NN algorithm. Outlier detection is used for identification of items, events or observations which do not conform to an expected pattern or other items in dataset. The identification of instances that diverge from the expected behavior is important task. Existing techniques provides a solution to the problem of anomaly detection in categorical data with a semi supervised setting. The outlier detection approach is based on distance learning for categorical attributes (DILCAs), a distance learning framework was introduced. The key intuition of DILCA is that the distance between the two values of a categorical attribute can be determined by the way, in which they co-occur with the values of other attributes in the data set. In this outlier detection algorithm is work well for fixed-schema data, with low dimensionality. This work proposes novel anonymization methods for sparse high-dimensional data. It is based on approximate Classic K-Nearest Neighbor search in high-dimensional spaces. These representations facilitate the formation of anonymized groups with low information loss, through an efficient linear-time heuristic. Among the proposed techniques, Classic KNN-search yields superior data utility, but incurs higher computational overhead. In addition dimensionality reduction technique is used. In this work healthcare dataset are used. From the dataset 1000 instances are taken into consideration for providing enhanced outlier detection using Classic K-NN algorithm.

Keywords— Anomaly detection, categorical data, classic KNN approach, distance learning, semi-supervised learning.

I. INTRODUCTION

Enhanced Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning and supervised learning. Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy.

SEMI-SUPERVISED CLASSIFICATION

SSL (Semi-Supervised Learning) is a learning paradigm concerned with the design of models in the presence of both labeled and unlabeled data.

Depending on the main objective of the methods, this can divide SSL into a) Semi-Supervised Classification and b) Semi-Supervised Clustering.

Semi-Supervised Classification (SSC): It focuses on enhancing supervised classification by minimizing errors in the labeled examples, but it must also be compatible with the input distribution of unlabeled instances. Semi-Supervised Clustering: Also known as constrained clustering, it aims to obtain better-defined clusters than those obtained from unlabeled data.

ANOMALY DETECTION

In data mining, anomaly detection (also outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set, and then testing the likelihood of a test instance to be generated by the learnt model.

II. RELATED WORK

A. A semi-supervised approach to the detection and characterization of outliers in categorical data[2008]

This described distance learning for categorical attributes (DILCAs), a distance learning framework was introduced. The key intuition of DILCA is that the distance between the two values of a categorical attribute Ai can be determined by the way, in which they co-occur with the values of other attributes in the data set [1]. The added value of this proximity definition is that it takes into consideration the context of the categorical attribute, defined as the set of other attributes that are relevant and non-redundant for the determination of the categorical values. Relevancy and

redundancy are determined by the symmetric uncertainty (SU) measure that is shown to be a good estimate of the correlation between attributes.

B. Anomaly detection: a survey [2009]

This survey described the important problem that has been researched within diverse research areas and application domains [2]. Many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic. This survey tries to provide a structured and comprehensive overview of the research on anomaly detection. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effectiveness of the technique in that domain.

C. From context to distance: learning dissimilarity for categorical data clustering [2012],

This research described clustering by categorical attributes is a challenging task in data mining applications. Unlike numerical attributes, it is difficult to define a distance between pairs of values of a categorical attribute, since the values are not ordered. The authors propose [3] a framework to learn a context-based distance for categorical attributes. The key intuition of this work is that the distance between two values of a categorical attribute. They validate their approach by embedding their distance learning framework in a hierarchical clustering algorithm.

III.SEMISUPEVISED APPROACH IN CATEGORICAL DATA

The system methodology design an anomaly detection framework for categorical data based on the distance learning approach and embeds the distance learning algorithm within different ranking strategies. The proposed methodology gives a solution to the problem of anomaly detection in categorical data with a semi-supervised setting. Our approach is based on distance learning for categorical attributes (DILCAs).

The proposed unsupervised method for categorical data that marks as anomalies those instances whose compression cost is higher than the cost required by the norm in a pattern-based compression mechanism based on the minimum description length principle. The characterization is successively employed to detect the anomalous instances in a semi- supervised scenario.

This research proposes the distance learning approach has been successfully employed in a classification scenario. This project define a semi supervised anomaly detection framework for categorical data which takes the benefit of DILCA. Typically, in anomaly detection, there are two ways to present the results. The first one is to assign a normal/abnormal label to each test data instance. The second is to give an anomaly score (a sort of anomaly degree) to each tested instance.

Distance learning for categorical attributes

The distribution of the values in the contingency table may help to define a distance between the values of a categorical attribute, but also that the context matters. Let us now consider the set F = { X1, X2,. . . , Xm} of m categorical attributes and data set D, in which the instances are defined over F. We denote by Y ∈ F the target attribute, which is a specific attribute in F that is the target of the method, i.e., the attribute on whose values we compute the distances.

It allows to compute a context-based distance between any pair of values (yi, y j) of the target attribute Y on the basis of the similarity between the probability distributions of yi and y j given the context attributes, called C(Y) ⊆ F \ Y.

XFor each context attribute Xi ∈ C(Y), computes the conditional probability for both the values yi and y j given the values xk ∈ Xi, and then, it applies the Euclidean distance. The Euclidean distance is normalized by the total number of considered values as

$$d(y_i, y_j) = \sqrt{\frac{\sum_{X \in \mathcal{C}(Y)} \sum_{x_k \in X} (P(y_i|x_k) - P(y_j|x_k))^2}{\sum_{X \in \mathcal{C}(Y)} |X|}}$$

At the end of the process, returns a distance model M = {M_{Xi} | i = 1,. . ., m}, where each M_{Xi} is the matrix containing the distances between any pair of values of attribute Xi.

Semi-supervised Anomaly Detection

In anomaly detection is implemented in two ways. The first one is to assign a normal/abnormal label to each test data instance. The second is give an anomaly score to each tested instance.

Depending on the constraints with respect to the admitted false positives or true negatives present in the results, the user may set a high or low threshold, or decide to consider a high or low value of k. Our approach supplies the second type of output: given a training data set D, the normality model learned on D, and a test instance t ∈ T, it returns the value of the anomaly score of t.

It learns a model consisting of a set of matrices M = {MXi }, one for each attribute Xi ∈ F. Each element mi(j, l) = d(xi j, xli) is the distance between the values x I j and xli of the attribute Xi , computed using distance learning for categorical attributes by the evaluation over the training data set D.

These matrices provide a summarization in terms of the distance learning for categorical attributes distance function on the distribution of the values of the attributes Xi given the other attributes in the instances of the normal class. From matrices MXi ,it is possible to compute a distance between any two data instances d1 and d2 on the basis of the DILCA distance between the categorical values, using the following formula:

$$\text{dist}(d_1, d_2) = \sqrt{\sum_{M_{X_i} \in \mathcal{M}} m^i(d_1[X_i], d_2[X_i])^2}$$

$$\text{OS}(t) = \sum_{p=1}^{k} \text{dist}(t, d_p)$$

Finally, measures the outlier score (OS) associated with each test instance t ∈ T as the sum of the distances between t and a subset of k ($1 \leq p \leq k \leq n$) instances d p belonging to D.

IV CALSSIC K-NEAREST NEIGHBOR SEARCH

The Classic k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification.

In this process, the numerical columns values of patient profiles (which are classified into specified groups already) is taken as x-axis and y-axis data and then new patient is classified into one of the existing patients' classes.

CLASSIC K-NEAREST NEIGHBORS ALGORITHM

The Classic k-Nearest Neighbors algorithm (or Classic k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression

- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.
- If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class or the object property value is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

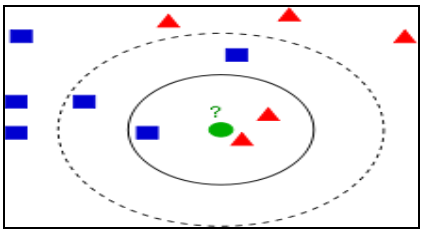

Fig 4.1 Classic k-NN classification

Data Reduction

Data reduction is one of the most important problems for work with huge data sets. Usually, only some of the data points are needed for accurate classification. Those data are called the prototypes and can be found as follows:

1. Select the class attribute-outliers, that is, training data that are classified incorrectly by k-NN (for a given k)

2. Separate the rest of the data into two sets: the prototypes that are used for the classification decisions and the absorbed points that can be correctly classified by k-NN using prototypes. The absorbed points can then be removed from the training set.

Selection of class outliers

A training example surrounded by examples of other classes is called a class outlier.

- Random error calculated (measurements are caused by unknown and unpredictable changes).

- Insufficient training examples of this class (an isolated example appears instead of a cluster)

- Missing important features (the classes are separated in other dimensions which we do not know)

- Too many training examples of other classes (unbalanced classes) that create a "hostile" background for the given small class

Class outliers with k-NN produce noise. They can be detected and separated for future analysis. Given two natural numbers, k>r>0, a training example is called a (k, r) NN class-outlier if its k nearest neighbors include more than r examples of other classes.

Classic KNN for data reduction

Classic K-Nearest Neighbor (KNN, the Hart algorithm) is an algorithm designed to reduce the data set for k-NN classification. It selects the set of prototypes U from the training data, such that 1st NN with U can classify

the examples almost as accurately as 1st NN does with the whole data set. Given a training set X, KNN works iteratively:

- Scan all elements of X, looking for an element x whose nearest prototype from U has a different label than x.
- Remove x from X and add it to U
- Repeat the scan until no more prototypes are added to U.

Use U instead of X for classification. The examples that are not prototypes are called "absorbed" points. It is efficient to scan the training examples in order of decreasing border ratio. The border ratio of a training example x is defined as

$$a(x) = ||x'-y|| / ||x-y||$$

Where ||x-y|| is the distance to the closest example y having a different color than x, and ||x'-y|| is the distance from y to its closest example x' with the same label as x.

The border ratio is in the interval [0,1] because ||x'-y|| never exceeds ||x-y||. This ordering gives preference to the borders of the classes for inclusion in the set of prototypes U. A point of a different label than x is called external to x. The calculation of the border ratio is illustrated by the figure on the right. The data points are labeled by colors: the initial point is x and its label is red. External points are blue and green. The closest to x external point is y. The closest to y red point is x'. The border ratio a(x)=||x'-y||/||x-y|| is the attribute of the initial point x. In k-NN regression, the k-NN algorithm is used for estimating continuous variables. One such algorithm uses a weighted average of the k nearest neighbors, weighted by the inverse of their distance.

This algorithm works as follows

- Compute the Euclidean is distance from the query example to the labeled examples.
- Order the labeled examples by increasing distance.
- Find a heuristically optimal number k of nearest neighbors, based on RMSE. This is done using cross validation.
- Calculate an inverse distance weighted average with the k-nearest multivariate neighbors.

Dimensionality reduction for KNN search

In the dimensionality reduction for KNN search, the three numerical columns values of patient profiles (which are classified into specified groups already) are taken and two of the three data are averaged into one data and is taken as x-axis and third column as y-axis data and then new patient is classified into one of the existing patients' classes.

V EXPERIMENTAL ANALYSIS

Experimental analysis is to be of use to researchers from all fields who want to study algorithms experimentally. To demonstrate the proposed method classic KNN classification is used and compare it performance with KNN Classification.

The following Table 5.1 describes Secure Outlier model for existing KNN and Classic KNN classification algorithm. The table contains number of patient datasets, average for KNN Classification algorithm and average performances for Classic-KNN Classification algorithm details are given below.

Number of Datasets [N]	KNN-Classification [%]	Classic KNN-Classification [%]
100	50.2	51.33
200	58.67	59.32
300	64.03	65.34
400	72.33	73.44
500	76.12	77.98
600	79.33	79.89
700	80.44	81.04
800	81.45	82.78
900	83.22	84.03
1000	84.10	85.65

Table 5.1 Performances Analysis- KNN –Classic KNN Classification Algorithm

The following Fig 5.1 describes Secure Outlier model for existing KNN and Classic KNN classification algorithm. The figure contains number of patient datasets, average for KNN Classification algorithm and average performances for Classic-KNN Classification algorithm details are given below.

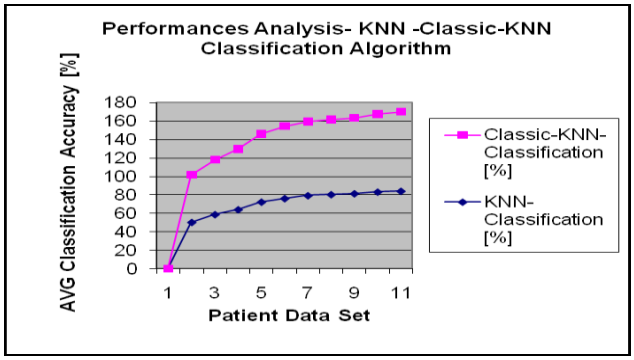

Fig 5.1 Performances Analysis- KNN-Classic-KNN Algorithm

The following Table 5.2 describes Secure Outlier model for existing KNN and Classic KNN Classification algorithm. The table contains number of observation patient dataset, average for KNN Classification algorithm and average performances for Classic-KNN Classification algorithm details are given below.

Number of Numerical Datasets [N]	KNN Classification [%]	Classic KNN Classification [%]
100	60.43	58.23
200	63.66	61.67
300	65.04	62.22
400	67.99	64.83
500	69.03	66.45
600	71.95	68.80
700	72.77	69.34
800	74.03	70.07
900	75.88	72.09
1000	76.73	73.08

Table 5.2-Performance Analysis of K-NN & Classic-KNN Classification Algorithms

The following Fig 5.2 describes Secure Outlier model for existing KNN and Classic KNN Classification algorithm. The figure contains number of observation patient dataset, average for KNN Classification algorithm and average performances for Classic-KNN Classification algorithm details are given below.

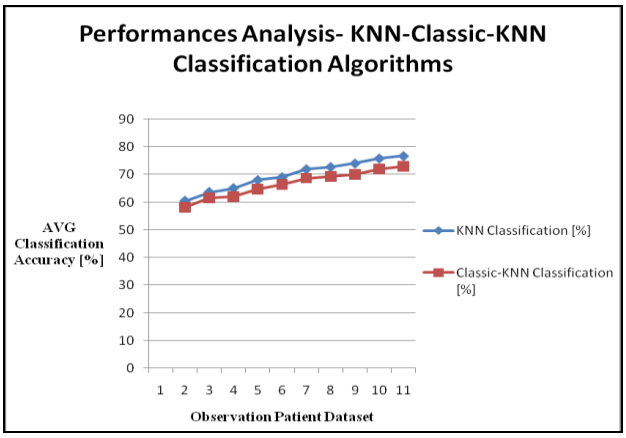

Fig 5.2: Performance Analysis of K-NN and Classic-KNN Classification Algorithm

The following Table 5.3 describes Secure Outlier model for existing KNN and Classic KNN Classification algorithm. The table contains number of observation patient dataset, time taken for KNN Classification algorithm and time taken performances for Classic-KNN Classification algorithm details are given below.

Number of Numerical Datasets [N]	KNN Classification [ms]	Classic KNN Classification [ms]
100	0.008	0.005
200	0.014	0.011
300	0.023	0.020
400	0.032	0.026
500	0.042	0.037
600	0.053	0.044
700	0.065	0.052
800	0.072	0.064
900	0.081	0.075
1000	0.094	0.083

Table 5.3-Performance Time Analysis of KNN and Classic-KNN Classification

The following Table 5.3 describes Secure Outlier model for existing KNN and Classic KNN Classification algorithm. The figure contains number of observation patient dataset, time taken for KNN Classification algorithm and time taken performances for Classic-KNN Classification algorithm details are given below

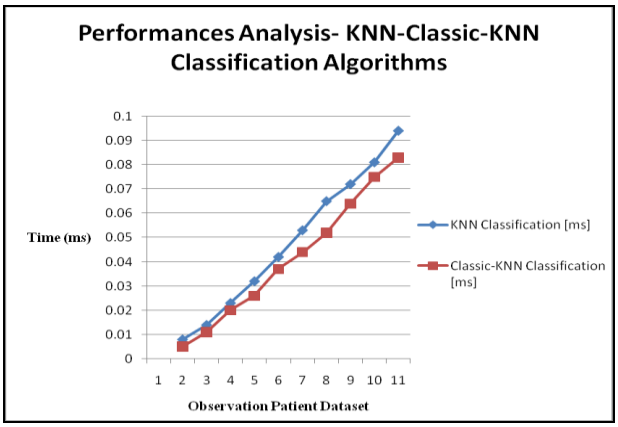

Fig 5.3 Performance Time Analysis of KNN-Classic-KNN Classification

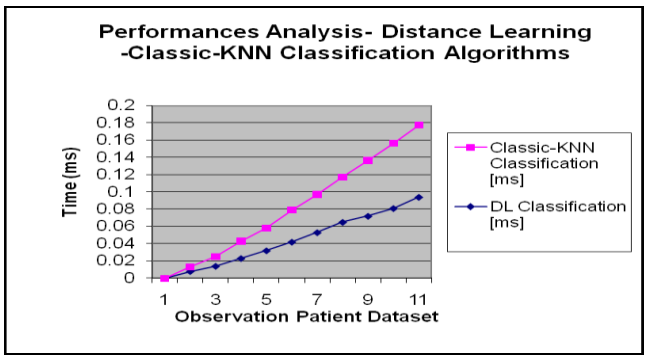

Fig 5.4 Performance Time Analysis of DL-Classic-KNN Classification

VI CONCLUSION AND FUTURE ENHANCEMENT

Outlier detection is an important issue occurs within various research and applications domains in today. It aims to detect the object that are considerably distinct, exceptional and inconsistent the majority data in input data sets. In this research study identify abnormal data which forms non-conforming pattern is referred to as outlier, anomaly detection. This leads to knowledge and discovery.

In this research an empirical system is developed to test the proposed methodology with the three numerical columns values of patient profiles (which are classified into specified groups already) are taken and two of the three data are averaged into one data and is taken as x-axis and third column as y-axis data and then new patient is classified into one of the existing patients' classes.

To reduce redundant or irrelevant features that can improve classification performance in most of cases and decrease cost of classification. A novel semi-supervised classification technique has to be proposed for dimensionality reduction in mammogram classification. As a future work, new data structures to handle categorical data more efficiently and speed up the anomaly detection task for continuous and categorical attributes.

ACKNOWLEDGMENT

REFERENCES

1. Dino Ienco, Ruggero G. Pensa, and Rosa Meo," A Semisupervised Approach to the Detection and Characterization of Outliers in Categorical Data", IEEE transactions on neural networks and learning systems, February 2, 2016.
2. Chandola .V, Banerjee .A, and Kumar .V, "Anomaly detection: A survey," ACM Compute Survey., vol. 41, no. 3, 2009, Art. ID 15.
3. Ienco .D, Pensa R.G, and Meo .R, "From context to distance: Learning dissimilarity for categorical data clustering," ACM Trans. Knowl. Discovery Data, vol. 6, no. 1, 2012.
4. A. Ahmad and L. Dey. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. Pattern Recogn. Lett., 28(1):110–118, 2007.
5. A. Koufakou, E. G. Ortiz, M. Georgiopoulos, G. C. Anagnostopoulos, and K. M. Reynolds, "A scalable and efficient outlier detection strategy for categorical data," in Proc. 19th IEEE ICTAI, Patras, Greece, Oct. 2007, pp. 210–217.
6. Akoglu .L, Tong .H, Vreeken .J, and Faloutsos .C, "Fast and reliable anomaly detection in categorical data," in Proc. 21st ACM CIKM, Maui, HI, USA, Oct. 2012, pp. 415–424.
7. Angiulli .F and Fassetti .F, "Distance-based outlier queries in data streams: The novel task and algorithms," Data Mining Knowl. Discovery, vol. 20, no. 2, pp. 290–324, 2010.
8. Breunig M. M, Kriegel H.P, Ng .R.T, and Sander .J, "LOF: Identifying density-based local outliers," in Proc. ACM SIGMOD Conf. Manage. Data, Dallas, TX, USA, May 2000, pp. 93–104.
9. Chandola .V, Boriah .S, and Kumar .V, "A framework for exploring categorical data," in Proc. SIAM Int. Conf. Data Mining, Sparks, NV, USA, Apr. 2009, pp. 187–198.
10. Hido .S, suboi Y.T, Kashima .H, Sugiyama .M, and Kanamori .T, "Statistical outlier detection using direct density ratio estimation," Knowl. Inf. Syst., vol. 26, no. 2, pp. 309–336, 2011.
11. Hido, S., suboi, Y.T, Kashima, H., Sugiyama, M. and Kanamori, T., Inlier-based outlier detection via direct density ratio estimation, in 'Proceedings of the 8th IEEE International Conference on Data Mining', pp. 223–232.
12. K. Das, J. Schneider, and D. B. Neill, "Anomaly pattern detection in categorical datasets," in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Las Vegas, NV, USA, Aug. 2008, pp. 169–176.
13. K. Smets and J. Vreeken, "The odd one out: Identifying and characterising anomalies," in Proc. SIAM Int. Conf. Data Mining, Mesa, AZ, USA, Aug.2011, pp. 804–815.
14. **M. Lichman.** (2013). UCI Machine Learning Repository.[Online].Available: http://archive.ics.uci.edu/ml.
15. **Noto .K, Brodley .C, and Slonim .D,** "FRaC: A feature-modeling approach for semi-supervised and unsupervised anomaly detection," Data Mining Knowl. Discovery, vol. 25, no. 1, pp. 109–133, 2012.