Abstract — The outlier detection in the field of data mining and Knowledge Discovering from Data (KDD) is capturing special interest due to its benefits. It can be applied in the financial area, because the obtained data patterns can help finding possible frauds and user errors. Therefore, it is essential to assess the truthfulness of the information. In this context, data auditory process uses techniques of data mining that play a significant role in the detection of unusual behavior. Here, a method for detecting values that can be considered as outliers in a nominal database is proposed. The basic idea in this method is to implement: a Global k-Nearest Neighbors algorithm, a clustering algorithm named k-means, and a statistical method of chi-square. The application of algorithms has been developed with a database of candidate people for the granting of a loan. Each test was made on a dataset of 1180 registers in which outliers have been introduced deliberately. The experimental results show that the method is able to detect all introduced values, which were previously labeled to be differentiated. Consequently, there were found a total of 48 tuples with outliers of 11 nominal columns.

Keywords — outlier, data mining, KNN, chi-square, financial fraud

I. INTRODUCTION

Detecting outlier represents a challenge in data mining techniques. Outliers have different characteristics comparing to other data stored in a database; due to their behavior, they are not similar to the majority. Also, they are susceptible to be introduced for malicious mechanisms [1]. Moreover, Mandhare et al., [2] consider this data type as a threat, defining it as irrelevant or malicious, generating conflicts during the analysis process and generating unreliable and inconsistent results.

The detection of outliers facilitates the finding of anomalous values or unusual patterns in data. Several studies show that most of this data type is originated from domains such as credit cards [3], security systems [4], and e-health information [5].

Also, detection in data mining activities uses tools based on non-supervised algorithms [6]. Those activities have a local or global scope [7]. On the one hand, global approaches are referred to as techniques in which it is assigned punctuation for each anomaly in each instance respect to the global dataset. On the other hand, local approaches represent the anomaly in a point of data with respect to their direct neighborhood. Local approaches detect atypical values that are ignored when global approaches are used, especially in variable density [8]. Examples of said algorithms are those based on i) clusters and ii) the nearest neighbor. The algorithm of the former category considers that the atypical values are in disperse neighborhoods, which are far from the nearer neighbors. While the second category operates in the grouped algorithms [6].

Besides, other studies are focused on the outliers register detection and not only on the detection at column level [2], [9]–[14]. Kuna et al., [15] present a procedure that applies decision trees based on the C4.5 algorithm, which uses continue values and separates the possible results on two branches. Then, a decision tree is generated with these data by means of partitions obtained recursively. To find the more significative attributes and then each input and output attribute, it is necessary to apply the Local Outlier Factor (LOF). If the methodology is effective, the new algorithms need to be evaluated, based on the concepts of the “nearest neighbors” in order to improve the accuracy of the solution. There are studies that demonstrate up to 75% of improvement in the performance of the classification of this algorithm [16]. In addition, there are studies that present an adaptive and cluster-based hybrid method to accelerate this classification algorithm taking into account the response time in large datasets [17].

This paper proposes a methodology for the detection of atypical values, which are based on the application of traditional statistical methods (i.e., chi-square test) in synchrony with data mining algorithms (i.e., KNN Global Anomaly and K-means). Also, the method has been applied to the financial domain. Finally, an evaluation has been performed, where good insights about the application of the methodology and its performance in the treatment of outlier have been obtained, even when there are null values.

This study is structured as follows: Section 2 discusses the existing solutions. Section 3 presents the proposed method. Section 4 presents the evaluation of the application of the methodology, and finally, Section 5 presents the conclusions and further work.

II. RELATED WORK

There are several approaches related to outlier detection, in this context, Patel et al., [18] have performed a survey where it is presented a summary of the different outlier detection studies, which are statistic-based approach, distance-based approach, density-based approach. The authors contribute with a discussion related to an outlier, the methods used for clustering the dataset, and finally, they conclude that k-mean algorithm is the most widely used for clustering a dataset. On the other hand, there are studies such as [2], [9]–[14], those
approaches include the use of data mining techniques, statistical methods, or both. For the outlier detection, commonly the nearest-neighbor techniques have been used (i.e., KNN) along with other techniques in order to find unusual patterns during the data behavior or to improve the performance of the process. Ying Gu et al., [13] present an efficient grid-based method to find outlier data patterns in big datasets. Similarly, Ke Yan et al., [9] propose an outlier detection method with KNN and data prune. That method takes successive samples of tuples and columns and applies an algorithm KNN in order to reduce the dimensionality without losing relevant information. In relation to the use of data mining techniques applied to the financial sector, the study presented by Malini and Pushpa [12] explains the outlier detection for credit cards frauds with techniques that detect unusual patterns through the KNN algorithm. The technique detects the input transaction and calculates the nearest point and the similarity score, determining if the transaction is a fraud.

A similar study is presented by Venkata et al., [10] which applies KNN by using a technique that uses a window with assigned memory, in that way, when it is detected a request transaction, it needs only one scan of the current window to find the object whose k-nearest neighbors are influenced.

These approaches use the KNN technique with the measures of traditional distance and similarly to detect fraudulent transactions; however, this paper uses a KNN variation named KNN Global Anomaly, which uses directly an anomaly measure, which is more useful for this research.

This research includes a method that uses the chi-square statistical method to determine and select the most relevant columns in the study. A related study has been developed for Sumaiya [11]; where the authors use the chi-square during the intrusion detection into a computer network to segment it into normal and abnormal traffic. The process includes the use of the Support Vector Machine (SVM) technique. However, there are no studies that combine the statistical method chi-square with KNN; which can help to pre-categorize the important columns before applying the KNN technique.

III. A HYBRID METHODOLOGY FOR OUTLIERS FINDING

In this section, it is presented a methodology for finding atypical values into a nominal database (dataset). This proposal evaluates records and columns (customer profile) with respect to the output column (credit amount granted) to find outliers values. The output column named S is a discretized value that represents the values obtained based on a set of input columns.

The method of finding outliers values is divided into three phases of development: i) the first phase uses the chi-square statistical method to define a weight W, which indicates the relevance of the client's profile compared to the amount of credit granted, ii) in the following phase, it has been used a grouping algorithm named KNN Global Anomaly (KNN) which calculates the score of outliers values, and iii) K-Means algorithms is used in order to separate outliers false-positive values.

A. Proposed Methodology

In this subsection the proposed methodology is detailed. Here, it has been used the Software Process Engineering Metamodel (SPEM 2.0). Figure 1 shows the activities applied to find outliers in which, three important phases are presented; which are conveniently numbered in Figure 1 and described below.

1. Classifying the most significant columns in relation to the output column through the Chi-square statistical method.
2. Qualifying the nearest global neighbor to obtain the values with suspected abnormality through the algorithm (KNN Global Anomaly).
3. Group the values with suspicion of an anomaly in outliers and false-positives through the application of the K-means grouping algorithm.

For a better understanding, the applied techniques which are used as guidelines in SPEM are described:

1) Chi-Square Statistical Tool: It is a nonparametric technique used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies [19]. The weight of the input columns (customer profile) was calculated related to the output column (credit amount). The greater the weight of a column on a scale of 0 to 1, the more relevant it is considered. The method can only be applied to columns of nominal type and it was selected as a method to define relays instead of using average or variances, this test uses frequencies [19].

The proposed methodology was used to perform a hypothesis test on the weight or importance of each of the columns with respect to the column S. Chi-square reports the level of significance of the associations or dependencies of each column of the client's profile compared to the output value S. This value is stored in a column called weight, which, together with the anomaly score, are reported in the final report. This column will serve as a reference for the decision making of the stakeholders.

2) The KNN Global Anomaly Score: The operator KNN Global Anomaly Score or KNN is based on the closest neighbor k algorithm, which calculates the anomaly score of the data in relation to the neighborhood [20]. Usually, outliers are far from their neighbors or the neighborhood is scarce. In the former case, it is known as global anomaly detection and it is identified with KNN; the second case refers to an approach based on local density. The score comes from the average of the distance to the nearest neighbors [8]. In the classification of the nearest neighbor k, the output column S is related to a new data not classified in the prediction, this implies a line of linear decision.

To obtain a correct prediction, the value k must be carefully configured. A high value of k represents a bad solution with respect to the prediction, while low values
generate noise [21]. Frequently, the parameter $k$ is chosen empirically and depends on each problem. Hassanat, Abbadi, Altarawneh, & Alhasanat (2014) propose testing with different numbers of close neighbors until they reach the one with the best accuracy. Their proposal starts with values from $k = 1$ to $k = \sqrt{n}$, where $n$ identifies the number of records in the training dataset. The general rule is often to assign $k$ with the square root of the number of records in dataset $D$.

Taking into account the previous considerations, $k = 34$ was configured for the dataset (see equation 1); the resulting value uses the equation, where $n$ identifies the number of records in the dataset and $\text{TRUNC}$ to the function of truncation of the value obtained from the square root.

$$k = \text{TRUNC}(\sqrt{n}) \quad (1)$$

The KNN process includes a data preparation procedure, which converts the analysis column into a record. That is, the name of the column is part of the record in a new column called attribute. The execution of the algorithm KNN generated a column called outlier; which records the weight of the anomaly of the data in relation to all the records in dataset $D$. The result of the outlier column varies according to the distance of each of the data with respect to the average distance of the nearest neighbors. The records that have values higher than 0 are those that are likely to be classified as outliers values; however, due to the variability of values obtained in each iteration, and to maintain uniformity in the selection criteria for the columns examined, the values were normalized on a scale between 0 and 1. The value 0 corresponds to the records that are not likely to be outliers. If the outlier is greater than 0 and tends to 1, it means that the record is more likely to be an outlier. The value of the outlier column was included in the final report for further detailed analysis.

The accuracy of a KNN algorithm is not always optimal in the analysis of all columns due to cardinality; which is a relevant characteristic that needs to be taken into account; thus, each column varies in direct relation to the information. That is, in the case of discrete values with low cardinality, the accuracy increases, due to the fact that the frequency is concentrated in few values. As a result, values not classified in these concentrations were considered as outliers. For example: if there is a column named "genre", it admits a cardinality of ("M", "F", "O"). The values that did not coincide with these two groups were classified by KNN with a score of 1 in the outlier column and consequently classified as outliers. The next case represents the values with a high cardinality, where many columns with smaller frequencies are formed. In this case, the KNN algorithm decreases its accuracy in classes separation (i.e., normal, abnormal) due to the high dispersion. For example, in an age column, it has a high cardinality due to the high number of classes that are created.

3) The $K$-means Algorithm: It is one of the most popular grouping algorithms. The algorithm is based on distances and is used to divide the data into a given number of clusters [22]. The data grouped in the same cluster are similar compared to other clusters. K-means randomly assign a point called centroid which is based on the value of the parameter $k$, from which distances are taken to the other points. The distance metric can be euclidean, cosine or fast cosine distance (Euclidean distance was used for this case). Once the distance is calculated, each data is assigned iteratively to the corresponding cluster according to the similarity as a function of the distance and the average value of data. Next, the centroids and the average value of the distance to the data are recalculated until the centroids do not move [23].

The $K$-means algorithm guarantees the result of the KNN method. It is important to consider the high dispersion columns. It has been applied to an evaluation process addressed by the distance. This procedure divides the dataset of each column into two clusters: i) false-positives and ii) outliers. Therefore, it has been executed the $k$-means algorithm with the parameter $k=2$, it is shown that the algorithm divides the analyzed data into two groups. A performance measurement process

In this case, the $k$-means algorithm was executed with the parameter $k=2$, which indicates that the algorithm must form two groups of values in the analyzed data. A performance measurement process was necessary to generate a distance-vector consisting of global average distance, average distance to cluster 0 and average distance to cluster 1. Between the point that marks the average distance of the clusters and each cluster, there is a distance that separates the outliers from false-positives. The cluster with the greatest distance to the global average corresponds to the cluster of outliers. The calculations were made with an index called distance factor, in which scores range between 0 and 1. These values correspond to the proportionality of the distances, with the degree of distance between the clusters and the global average. That is, a significant relationship between the general average distance and the distance of the cluster 0, compared to the general average distance and the distance of the cluster 1, determined the cluster that has the outliers of those that are false-positive (see Figure 2).

At the beginning of the process, a score was assigned in each pair of columns. Here, one of the columns is named an input column; the output column is the objective. The values stored in the weight column are the product of the calculation of the chi-square statistic in a range of 0 to 1 and establish the relevance of each column with respect to the $S$ output. Values with a tendency to one are those most relevant in terms of their influence on column $S$. Once KNN and $K$-means were executed, a filter was established that limits the output report based on the weight of the anomaly. Initially, only those values that exceed a weight greater than 0 were placed.

IV. Evaluation and Results

In order to evaluate the method for detecting outliers, a dataset was generated in which values that represent abnormalities according to classification and frequency
criteria have been introduced. To include those values, the following steps have been performed.

1. Identifying enough number of characteristics to construct the classification criteria.
2. Determining according to the frequency, the metric to establish the division between the data.
3. Establishing a criterion based on the results of the frequencies of the observations, which are classified as outliers.

The inclusion of outliers was completed according to the aforementioned approach, identifying the records and the values of the columns with a suspected abnormality. The objective will be to make a comparison of this identification phase with the results of the procedure.

The dataset stated corresponds to a credit database of nominal type with 15 columns as is indicated in Table 1. A total of 13 of the columns correspond to the personal characteristics of the applicant (customer profile); another to the identifier of the account "Account Id", and finally "Credit granted" represents the discretized value of the credit amount. The latter is also called: exit or target column. The dataset comes from a demonstration used in test environments to find outliers. It has 1180 records and contains null values.

| Account Id | Status | Country |
|------------|--------|---------|
| Age        | Civil status | Anual_income |
| Gender     | Children | Education |
| Antiquity  | Credit Card | Job |
| Is_propietary | Credit_card_qty | Credit granted |

The software used for experimentation was Rapid Miner Studio in version 8.1.01. Additionally, the Anomaly Detection extension version 2.4.001 of German Research for Artificial Intelligence was installed.

The dataset was verified for consistency and truthfulness as a whole. It was observed that the Account Id contains repeated and altered values, for this a new alternate identifier was generated.

During the experimentation procedure, each column in the database was analyzed in terms of the information it contains, for this purpose frequency tables were implemented. The values with low frequency with respect to the global result had the main suspicion of belonging to the group of outlier values. The observation also examined values in relation to the patterns of context behavior; candidates for outlier were determined. The values obtained were investigated in each experiment to verify the effectiveness of the procedure.

B. Validation

In order to give reliable results, a validation process was carried out in which, a column-by-column database survey was previously carried out, grouping each of the values, not including repeated values. Each of the values is accompanied by the repetition frequency established by the first anomaly inquiry. For example, if the “Sex” column is analyzed; 605 of the values are "F", 574 are "M", and a value contains the value "4", it could be ensured that the outliers in the procedure should identify is the value "4" and not "M" that corresponds surely to masculine and "F" to feminine.

C. Results

The procedure detected 100% of introduced outliers and false-positive values of a database for granting financial credits.

The execution of the chi-square statistical method allowed defining the attributes that do not provide a degree of the decision in relation to the output column called “credit amount”. The column "country" does not define relevance, unlike the columns: "seniority", "owner_", "annual income" and "occupation", with weight values greater than 0.30 which determine a better significance with respect to columns with less assessment. The remaining columns with values greater than zero, although they do not have greater relevance, are considered within the procedure for the search for outlier values. Table 2 illustrates the values resulting from the Chi-square application.

| Attribute            | Weight |
|----------------------|--------|
| antiquity            | 1,00   |
| is_owner             | 0,92   |
| anual_income         | 0,65   |
| job                  | 0,32   |
| credit_card          | 0,16   |
| education            | 0,13   |
| civil_status         | 0,11   |
| credit_card_quantity | 0,04   |
| cta_number           | 0,02   |
| age                  | 0,02   |
| children             | 0,01   |
| status               | 0,01   |
| gender               | 0,01   |
| country              | 0,00   |

The analysis of outliers for the “Account Id” column was not considered in the study, since it firstly presents inconsistency in terms of coding with values of “-1” in 5 of the records, and second that the attribute is an identifier with high variability, not deterministic for values of the objective attribute.

Table 3 shows the results obtained in the application of the procedure with respect to the initial study of frequency tables.

A total of 48 tuples were found in 11 columns with outliers (some of the tuples can be duplicated by having more than one outlier column). Only the age column contains null values that were not considered in the analysis.

The second step of the procedure consisted in the application of the closest KNN algorithm on each column of the client profile. This was useful for obtaining an anomaly rating with respect to the dataset. The results of the application of the
algorithm located outlier values in 80% of the columns and a remaining 20% resulted in a mixture between outliers and false-positives values. The method consisted in the application of the algorithm with a parameter k = 2, which considers 2 clusters group of false-positive data, group of outliers. The difference with Kuna's work lies in the inclusion of a formula that measures the distances of each cluster to the point that determines the average total distances. Through this distance degree indicator, it was established that the cluster is composed of outliers. After refining the parameter of the degree of distance, atypical values of false-positives were separated in one hundred percent of cases.

The analysis of outliers values for the “Account Id” column was not considered in the study, since this presents inconsistency in terms of coding with values of “-1” in 5 of the records; and secondly, that column is an identifier with high variability and that is not determinant for the values of the output column.

A total of 48 records of 11 columns were found with outlier values (some of the columns can be duplicated by having more than one outlier value). Only the age column contains null values that were not considered in the analysis.

The qualification of the degree of relationship will be important for the auditor, who could relate it to the outlier, and establish a degree of influence on the final result. It is proposed as future work to carry out the tests with other data sets of various kinds to validate the effectiveness of the method and the efficiency in the use of computational resources. It is recommended to look for alternative techniques to improve the configuration of the variable k of the algorithm KNN in an automated way.

V. CONCLUSIONS

This paper aims at the generation of a hybrid method to discover outliers; also, as part of the evaluation, it analyzed a database for granting financial credits, composed by attributes that describe the client's profile. These data influence the output column, referring to the amount of credit granted to the client. The results of analyzing each of the attributes that provide insights about the client profile; also it showed the identification of those values that do not follow a typical behavior pattern; usually named outliers. The proposed procedure was feasible and adequate for the identification of outliers in a dataset with diversity in the variability of nonnal type data. The methodology has three important activities: the first milestone with a statistical method and two data mining techniques. Once the parameters and filters were configured in each method, and 100% of the outliers were detected. As further work, the combination of columns and values will be studied in order to determine patterns in relation to the output column; also, it is necessary to execute additional evaluation in which exists more datasets and different context to experiment with this solution.

ACKNOWLEDGEMENT

This research was supported by the vice-rectorate of investigations of the Universidad del Azuay. We thank our colleagues from Laboratorio de Investigación y Desarrollo en Informática (LIDI) at Universidad del Azuay who provided insight and expertise that greatly assisted this study. Part of this research is supported by the Design of architectures and interaction models for assisted living environments aimed at older adults project of the XVIII DIUC Call for Research.

REFERENCES

[1] A. C. Atkinson and D. M. Hawkins, “Identification of Outliers,” Biometrics, 1981.
[2] H. Mandhare and S. R. Idate, “Distance Based Outlier Detection and Density Based Outlier Detection Techniques,” in International Conference on Intelligent Computing and Control Systems ICICCS, 2017, pp. 931–935.
[3] R. Bansal, N. Gaur, and S. N. Singh, “Outlier Detection: Applications and techniques in Data Mining,” in Proceedings of the 2016 6th International Conference - Cloud System and Big Data Engineering, Confluence 2016, 2016.
[4] M. A. Khan, S. K. Pradhan, and H. Farima, “Applying Data Mining techniques in Cyber Crimes,” 2017 2nd Int. Conf. Anti-Cyber Crimes. ICACC 2017, pp. 213–216, 2017.
[5] H. Zhang and L. Wang, “An information-Theoretic outlier detection method for prescription data,” 2017 3rd IEEE Int. Conf.
[6] A. Onan, “A K-medoids based clustering scheme with an application to document clustering,” 2nd Int. Conf. Comput. Sci. Eng. UBMK 2017, pp. 354–359, 2017.

[7] P. M. Monamo, V. Marivate, and B. Twala, “A multifaceted approach to Bitcoin fraud detection: Global and local outliers,” Proc. - 2016 15th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2016, pp. 188–194, 2017.

[8] M. Amer and M. Goldstein, “Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner,” Proc. 3rd RapidMiner Community Meet. Conference (RCOMM 2012), 2012.

[9] K. Yan, X. You, X. Ji, G. Yin, and F. Yang, “A hybrid outlier detection method for health care big data,” Proc. - 2016 IEEE Int. Conf. Big Data Cloud Comput. BDCloud 2016, Soc. Comput. Networking, Soc. 2016 Sustain. Comput. Commun. Sustain. 2016, pp. 157–162, 2016.

[10] V. R. Ganji, “Credit card fraud detection using anti-k nearest neighbor algorithm,” Int. J. Comput. Sci. Eng., vol. 4, no. 6, pp. 1035–1039, 2012.

[11] I. Sumaiya Thaseen and C. Aswani Kumar, “Intrusion detection model using fusion of chi-square feature selection and multi class SVM,” J. King Saud Univ. - Comput. Inf. Sci., vol. 29, no. 4, pp. 462–472, 2017.

[12] N. Malini and M. Pushpa, “Analysis on credit card fraud identification techniques based on KNN and outlier detection,” Proc. 3rd IEEE Int. Conf. Adv. Electr. Electron. Information, Commun. Bio-Informatics, IEEEICB 2017, pp. 255–258, 2017.

[13] Y. Gu et al., “Grid-based outlier detection in large data sets for combine harvesters,” Proc. - 2017 IEEE 15th Int. Conf. Ind. Informatics, INDIN 2017, pp. 811–818, 2017.

[14] T. T. Dang, H. Y. T. Ngan, and W. Liu, “Distance-based k-nearest neighbors outlier detection method in large-scale traffic data,” Int. Conf. Digit. Signal Process. DSP, vol. 2015-Septe, pp. 507–510, 2015.

[15] H. Kuna et al., “Procedimiento de Explotación de Información para la Identificación de Campos anómalos en Base de Datos Alfanuméricas,” Rev. Latinoam. Ing. Softw., vol. 1, no. 3, pp. 102–106, 2013.

[16] P. D. Rosero-Montalvo et al., “Neighborhood criterion analysis for prototype selection applied in WSN data,” Proc. - 2017 Int. Conf. Inf. Syst. Comput. Sci. INCISCOS 2017, vol. 2017-Novem, pp. 128–132, 2018.

[17] S. Ougiaroglou, G. Evangelidis, and D. A. Dervos, “FHC: An adaptive fast hybrid method for k-NN classification,” Log. J. IGPL, vol. 23, no. 3, pp. 431–450, 2014.

[18] S. P. Patel, “a Survey of Outlier Detection in Data Mining,” Int. J. Adv. Eng. Res. Dev., vol. 3, no. 01, 2015.

[19] M. Gol and A. Abur, “A modified Chi-Squares test for improved bad data detection,” 2015 IEEE Eindhoven PowerTech, PowerTech 2015, no. 1, pp. 1–5, 2015.

[20] P. Mulak and N. Talhat, “Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset,” vol. 4, no. 7, pp. 2101–2104, 2015.

[21] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study,” Decis. Support Syst., vol. 50, no. 3, pp. 602–613, 2011.

[22] R. T. Aldahdooh and W. Ashour, “DIMK-means ‘Distance-based Initialization Method for K-means Clustering Algorithms,’” Int. J. Intell. Syst. Appl., 2013.

[23] D. Sinwar and V. S. Dhaka, “Outlier detection from multidimensional space using multilayer perceptron, RBF networks and pattern clustering techniques,” Conf. Proceeding - 2015 Int. Conf. Adv. Comput. Eng. Appl. ICACEA 2015, pp. 573–579, 2015.

[24] H. Kuna et al., “Anomalos En Bases De Datos,” 2000.