Middle-East Journal of Scientific Research 23 (8): 1832-1838, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.08.103

Virtuous Approach for Outlier Detection in Image with Defective Data Labels

V. Lokesh, E. Bhuvaneswari, R. Maheswaran, K. Suruthi, E. Saranya and A. Kumaresan

Computer Science and Engineering, SKP Engineering College, Tiruvannamala, Tamil Nadu, India

Abstract: The digital image is getting increased in wider range. The key feature to be noted is the presence of outliers in the raw digital image. An outlier is an oddity which is deviating far away among the neighbour pixels which makes the digital image a noisy image. In the previous method, likelihood value of pixels is applied to eliminate the outliers by classifying the pixels label as normal and outlier. However, the digital image was not fully recovered back from the damaged pixels due to the imperfect labelling which causes the image to get distorted and the original quality of the digital image get spoiled. To overcome such problem, we use K-Means sparse vector decomposition (K-SVD) method in order to detect the presence of the outlier in the digital image. Our proposed method detects the outlier virtuously by computing the likelihood values among the neighbour pixels of the image. This shows that the proposed methodology works with a better trade off in false detection rate than the sate of art scenarios.

Key words: Outliers · Imperfect labelling · Anomaly detection · Likelihood

INTRODUCTION

Digital image processing and data mining is gaining importance in our day to day applications due to the increase in usage of data in large volumes. During the process of extracting data from larger volumes, an additional attention is needed due to the existence of unwanted oddity among the original set of data. This may be caused due to system errors like measurement of signal form a sensor or due to manual errors. This feature attracts various learning algorithms to make sure that the data to be extracted should be without any sort of deviations. Data which gets deviated much from the normal data is said to be outliers. In an image the outliers are referred to be noise which makes the scratch the pixels of the image this leads to damage in the original image. The outlier images or noisy image are due A well know definition of outlier is sated as “An outlying observation, or "outlier," is one that appears to move away from marked one with other members of the example in which it occurs” [1]. The presence of such irregularity among the perfect data is a deviation which is to be eliminated. The outlier which is identified will be labelled and it will be compared with other neighbouring data to find out the remaining outliers. Sometimes, the labelled data is not found to be an outlier (outliers are imperfectly labelled which leads to an inconsistent state) which is a major drawback prevailing in figuring out the exact imperfect data, To avoid such an imperfection, likelihood values of the data objects are found and they are clustered together based on their neighbour likelihood values. The data which are clustered will be classified as outliers and non outliers. Digital images acquired through many electronic devices are commonly subjected to the contamination of impulse noise. Some of the probable causes of impulse noise include malfunctioning pixel sensors, faulty memory units, imperfections encountered in channel during transmission. The outlier detection has become more precious in various applications like credit card fraud detection, fault medical diagnosis report generation of a patient, defence surveillances, intrusion detection for cyber-security etc.. (Figure 1) is an example of outlier among a set of data object located in a group from which a single data object is located far among the other data objects. This oddity of data is marked as an outlier.

Related Works: Identification of outlier in a data set is a rigid process. Since usage of database is increasing, supplementary maintenance of such a massive data has become complicated. The information in the database will not be accurate and consistent. The oddity among the set of data have to be marked and removed. In this segment,

Corresponding Author: V. Lokesh, Computer Science and Engineering, SKP Engineering College, Tiruvannamala, Tamil Nadu, India.

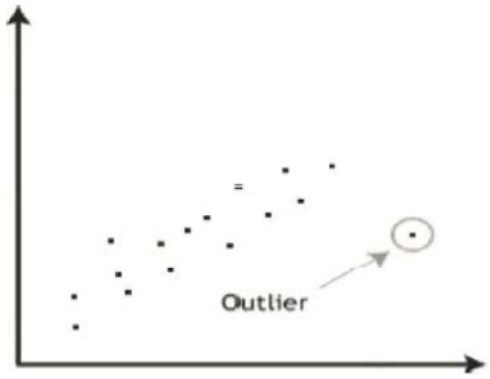

Fig. 1: A Single Outlier Deviates From the Group

the previous works so far carried out on outlier detection and its related works on other branches have been briefly reviewed. In data mining, outlier detection is the identification of objects, events or observations which do not match to a probable pattern or other items in data set [2]. Conventional outlier detection algorithms normally presume that outliers are hard or costly to obtain due to their unusual occurrences. The previous outlier detection algorithms are mostly classified into four categories. Statistics-based algorithms [2, 3]. Statistical approaches presume that the data follows some standard or fixed distributions and this kind of approach was used to find the outliers which deviate away from the distributions [4-6]. The methods in this type always assume the normal example follow certain kind of data distribution. However, we cannot always have this kind of priori data distribution knowledge to observe, particularly for high dimensional actual data sets [3]. For clustering-based approaches [7-9]; they always conduct clustering-based techniques on the samples of data to characterize the local data behavior. In general, the sub-clusters contain significantly a smaller amount of data points, after which additional clusters, are considered as outliers. For example, clustering techniques have been used to find anomaly in the intrusion detection domain [8]. In the work of [9], the clustering techniques iterative detect outliers to multidimensional data analysis in subspace. Since clustering based approaches are unsupervised without requiring any labelled training data, the performance is limited in unsupervised outlier detection. In density-based approaches a local outlier factor (LOF), [10-14] and variants [15] are the representatives of this kind of method. Depending on the local density of every data occurrence, the LOF determine the degree of outlierness, which provides suspicious status scores for all samples. The important property of the LOF is the capability to estimate local data structure by density estimation. The advantage of this method is that there will be no need to make an assumption on the generated distribution of the data. Conversely, these methods cannot sustain a high computational complexity in the testing period since they have to calculate the distance between all the other instances and each test instance to compute nearest neighbors. Besides the above work, model-based outlier detection approaches have been proposed [16-18]. Among them, support vector data description (SVDD) [16, 17] has been recognized empirically to be capable of identifying outliers in different domains. SVDD conducts a small sphere around the normal data and utilizes the constructed sphere to detect an unknown sample as normal or outlier. Using kernel function it can transform the original data into feature space in SVDD and effectively detect global outliers of high-dimensional data. Conversely, its performance is sensitive to the noise present in the input data.

One of the methods for evaluating outlierness is LOCI: Local Correlation Integral [19]. In this type, detecting outlier is highly effective than the previous methods because Loci are used to detect the outliers and group of outliers in a sample. A cut-off value will be generated automatically to conclude the point is an outlier. Based on the point value, the data are grouped into clusters and micro clusters to determine normal and outliers data values. Multi Granularity Deviation Factor (MGDF) [19] was introduced to detect the isolating outliers and to detect outlier in multi dimensional database where a deviation of event will occur rarely. This increase more attention on MGDF. The radius r for a point p_i at MGDF is the qualified deviation of its local neighbourhood density value from the average r-neighbourhood local density value. Therefore an object whose local neighbourhood density value matches the average neighbourhood local density will contain MDEF of 0.In distinction, outliers will contain MDEFs far from 0 from this spontaneous density values it is easy to conclude the normal objects from an outlier in a data set. LOCI has been extended as aLOCI approximate Local Correlation Integral which functions automatically to detect the outlier and micro cluster based on the cut off values and gives the accurate results. Fuzzy clustering is one of the methods used to detect one or more several outliers [20]. The method was used in both uvivariant and in multivariant.

Iii. Likelihood Values Generation: To find the outlier in a given training data set s this contains of l normal samples and a small quantity of n outlier samples.

Construct a classifier by means of normal and abnormal training data set and the classifier is applied to classify undetected test data. Due to occurrence of error during sampling or due to device imperfections [19], a normal sample may perform like an outlier. The example itself could not be an outlier. Such error can result in an improperly labelled training data, which makes the successive outlier detection disgustingly inaccurate. In order to avoid such problem, likelihood values between the data set was found and based on it the normal and abnormal data will be classified the following states the two likelihood model.

Single Likelihood Model: In this model, for each input data with a likelihood value will be associated $(x_i,m(x_i))$,which indicates degree of membership of an example towards its own class label.

Bi-Likelihood Model: In the model, all sample is associated with bi-likelihood values, denoted as $(x_i,m^t(x_i),m^n(x_i))$, In which $m^t(x_i)$ indicate the degree of an input data x_i belonging to the positive class and where $m^n(x_i)$ indicates the negative class respectively.

The goal is to compute likelihood values for every input data by means of creating a pseudo training dataset. For single likelihood model, the pseudo training data set consists of two parts specifically l normal samples and n abnormal samples are expressed as follows. $(x_1,m^t(x_1)m^n(x_1)),..,(x_1,m^t(x_1)),(x_{l+1},m^n(x_{l+1})),\ldots,(x_{l+n},m^t(x_{l+n}),m^n(x_{l+n}))$, in which $m^t(x_i)$ and $m^n(x_i)$ specify the likelihood of x_i belonging to the normal class and the abnormal, respectively.

For bilikelihood model the generated pseudo training data will be as follow:

$(x_1,m^t(x_1)m^n(x_1)),..,(x_1,m^t(x_1)m^n(x_1)),(x_{l+1},m^t(x_{l+1})m^n(x_{l+1})),\ldots,(x_{l+n},m^t(x_{l+n}),m^n(x_{l+n})),$

The basic idea of both schemes is to capture the local data uncertainty by examining the relation distances of every input data to its local neighbours in the feature space.

Kernel K-Means Clustering-Based Method: For each input data the kernel k-means clustering algorithm generate likelihood values.

$$J = \sum_{i=1}^{K}\sum_{j=1}^{l+n} \| \varnothing(x_j)-\varnothing(v_i)\|^2$$

where k represents the total number of clusters and v_i represents the cluster centre of the i^{th} cluster.k-means clustering generates set of local clusters, in which data samples of same cluster are more alike to each other, from this it is clear that if the majority of data samples belongs to same cluster are normal, it would contain a high probability chance of being normal. If there is any faraway point that does not belong to any cluster (outlying point), it would contain a high probability chance of being an outlier. For a given cluster j, presume there survive $l_j^p$of normal samples and l_j^n negative samples.

For the single likelihood model, the likelihood value of a normal example x_t specified to the normal class is calculated $m^t(x_t)l_j^p=l_j^p(l_j^p+l_j^n)$. Similarly, the likelihood value of an abnormal example x_k specifiedto the negative classis stated as $m^n(x_k) =l_j^n /(l_j^p+l_j^n)$.

For the bi-likelihood model, likelihood values of an example towards the normal and abnormal classes are calculated as $m^t(x_t) = l_j^p/ (l_j^p+l_j^n)$ and $m^t(x_t) = l_j^n /(l_j^p+l_j^n)$respectively.

Based on the kernel k-means clustering-method, the local data information of each sample is known. Based on the information, outliers are detected. The benefit of kernel k-means is that it aims to partition observations from the dataset into a group of local clusters but the major limitation is that it suffers well on datasets with changing densities which causes the inaccuracy over generated likelihood values to be[20].

Kernel LOF-Based Method: To contract with datasets with varying densities local density-based method has incorporated to compute likelihood values for each input data. The fundamental idea is to inspect the relation distance of a point in hyper sphere to its local neighbours in feature space. More specifically, for every point x_i, calculate its local reachability density first whose standard reachability distance determined based on the k-nearest neighbours of x_i.

$$Ir^{d_k}(x_i)= \frac{1}{K} \sum \text{reach-dist}_k(x_i, x_j)$$
$$xj \in N_k(x_i)$$

where $N_k(x_i)$ is the point $x_i$of k-nearest neighbors and the reachability distance of object x_j in the feature space to the object x_i is represented by reach-$dist_k(x_i,x_j)$. After finding the local reachability density next step is to find the nearest neighbors present in the feature space.

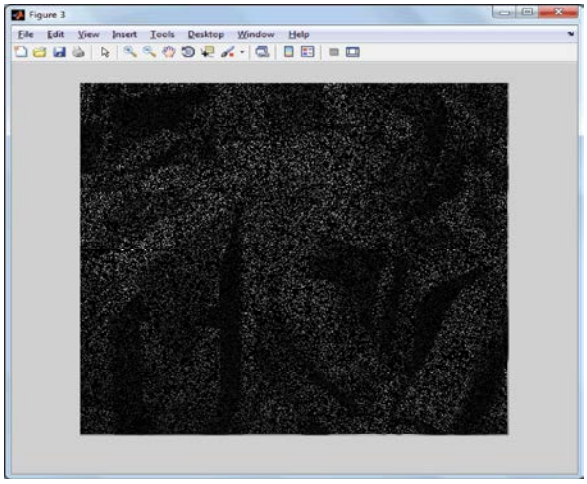

Fig. 2: Black And White Image with Outliers

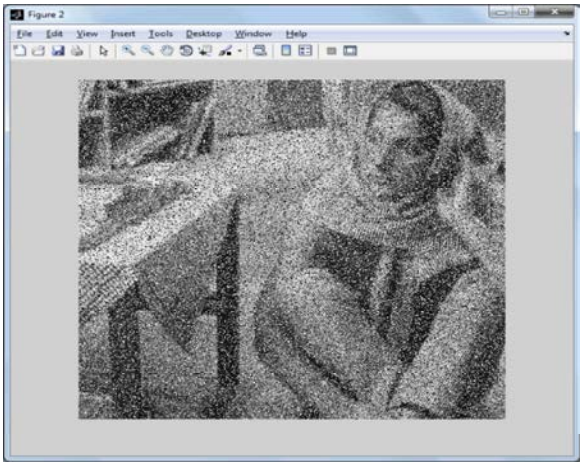

Fig. 3: Image after removal of Outliers

From this, the single likelihood and bi-likelihood values are calculated based on the local behavior of each sample this method helps to manage the dataset with varying densities [21].

Figure 2 represents a black and white image with more outliers which makes the raw image as noisy image. Figure 3 represent image after applying the existing approach, the image still remains with outlier and the quality of the image get damaged.

K-Means Sparse Vector Decomposition: K-SVD can be found commonly in many applications such as image processing, biology, audio processing and document analysis-Means sparse vector decomposition (K-SVD) is a generalization of the k-means clustering method and it is a singular value decomposition approach. K-SVD works by iteratively alternating between sparse coding the input data based on the current dictionary and updating the atoms in the dictionary to better fit the data.

$Y = \{y_i | i \in [1,K], y_i \in R^n\}$ as

$$\min_{D,X} = \|Y - DX\|^2_F \text{ subject to } \|x_i\|_0 < T \ \forall i \quad (1)$$

where X is formed by column stacking all vectors x_i and $\| \bullet \|^2_F$ denotes the frobenius norm square which is defined as the square of every element in the matrix.The K-SVD algorithm attempts to minimize the cost function iteratively, by first finding a coding for the signals This coding is sought such that it minimizes the error in representation and at the same time maintains a sparsity constraint. Once this sparse coding stage is done, the algorithm proceeds to update the atoms of the dictionary, one atom at a time, such that the error term is further reduced. Proceeding in such an iterative method, the algorithm reduces, or at worst maintains, the error of representation at each iteration. Having defined a general view of the steps of the algorithm, we proceed to detail the steps and the mathematical basis behind the method. We make an initial guess of the atoms of the dictionary which can be either from a set of over complete basis vectors or from the observed data itself. Given such an initial estimate D, the cost function of Eq. 1 can also be broken down into multiple optimization problems in the form

$$\min_{D,X} \|y_i - Dx_i\|^2 \text{ subject to } \|x_i\|_0 \ i = 1, 2, .., N \quad (2)$$

Now the problem of sparse coding is that of finding the code vector xi for each input signal y_i.Once an efficient coding vector is found for each signal, the K-SVD algorithm attempts to tune the dictionary D from the previous estimate, so as to further reduce the error. This is done one atom at a time. Thus at this stage of the algorithm, the matrix X is fixed and so is D with the exception of one column d_k under question. Let X_k^r denote the k-th row of the matrix X. The non-zero indices of X_k^r indicate all those signals which use the d_k atom for representation and the coefficient in the linear combination. For example, a value of 0.5 in the j-th index of X_k^r would signify that the 4 atom d_k is scaled 0.5 times and used as one of the atoms to represent signal y_j. The representation error term can thus be modified and written as

$$\|Y - DX\|^2_F = \|Y \sum_{J=1}^{K} d_j X_j^r \|^2_F$$
$$= \|Y - \sum_{J \neq k} d_j X_j^r - d_k X_k^r$$
$$= \|E_k - d_k X_k^r k\|^2_F.$$

this achieves is a separation of the error term into two parts – error when the atom dk is not taken into account and the error reduction due to its induction in reconstruction. This also achieves the decomposition of the multiplication of matrices into a summation of K rank-1 matrices. Of these, the first K - 1 are assumed to fixed. The problem of minimizing the total error thus boils down to finding a rank-1 matrix which best approximates the error matrix E_k. Estimation of such a matrix could easily be done by performing a singular value decomposition on E_k and using the largest singular value and its corresponding vector for this task. But such a solution will have no way of enforcing the sparsity constraint of the resulting X matrix. The authors propose a simple remedy to the problem. Instead of per-forming a straight SVD on the matrix E_k, a manipulation on the matrix is performed so as to bring it to a form where the SVD can be applied directly and hence implicitly maintain the sparseness of the result. To do this, we first need to identify all the signals that use the k-th atom of the dictionary. Once this is done the total error term of Eq. 3 can be split into two terms, one term defining the error of representation of those signals with the d_k atom removed and the rest for all other atoms. Eq. 3 thus takes the form of `

$$\|Y - DX\|^2_F = \|E_k^R - d_k x_k^R k\|^2_F \quad (4)$$

where E_k^R varies from E_{ke} of Eq. 3 only in the sense that it takes into account the error for just those signals that are supported by the atom d_k. The error function minimization can now be carried out by a rank- 1 approximation of the E_k^R matrix using singular value decomposition. The vector corresponding to the maximum singular value is then used. The interesting thing to note here is that this simple trick of selection of a subset of all the signals to reduce the error results in implicitly enforcing the sparsity constraint on the coding coefficients [22]. The reason for this is that only a single atom coefficient is updated at a time, that too only for those signals which base their representation on the dictionary atom d_k in question. Once the dictionary D is updated, one atom at a time, the algorithm again seeks to minimize the error in representation by recalculation of the coefficient matrix. The convergence of the algorithm is dependent on the

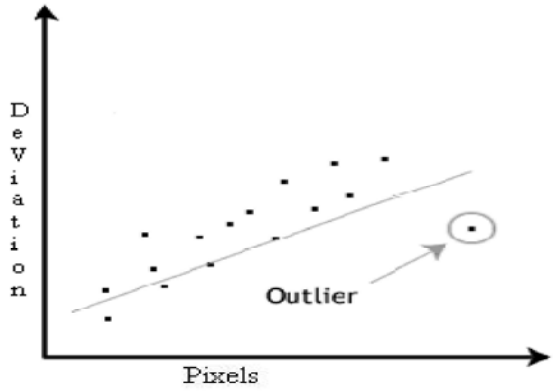

Fig. 4: Outlier detection among the image pixels

sparse coding function that is used to derive and approximate representation of each signal. However, if the threshold T is small enough, the solution of the sparse coding stage is close enough to the actually solution. Other more involved processes for sparse coding have been used by the OMP with all its simplicity delivers an acceptable result. Every such coding solution is hence guaranteed to reduce or keep unchanged the error in representation. A similar such statement can also be made for the dictionary updation process and hence it can be argued that with every iteration of the process, the cost function is minimized or remains the same. However, this does not guarantee convergence as it may well be possible to get stuck at a stable point [22].

Experimental Result: Outlier detection is a primary task in image processing and data mining to expose unusual patterns among the various domains our experimental research is focused on monitoring and apprehension of outlier behaviors of crowd in an image. The outlier's present in the image make the original images as noisy image to overcome such problem we use KSVD algorithm. The raw image with an outlier will be incorporated and the image is converted into an black and white image then the pixels are extracted from the black and white image and its local likelihood values are calculated. Pixels of image were extracted and it is taken as 3 X 3 matrix format. The Figure 4 represents the pixels present at the middle will be comparing with the nearby neighbour pixels values to detect the outliers presents in the image by finding the deviations. This process continues until each and every single pixels of the image get compared. The data values will be then categories as normal data and outlier data.

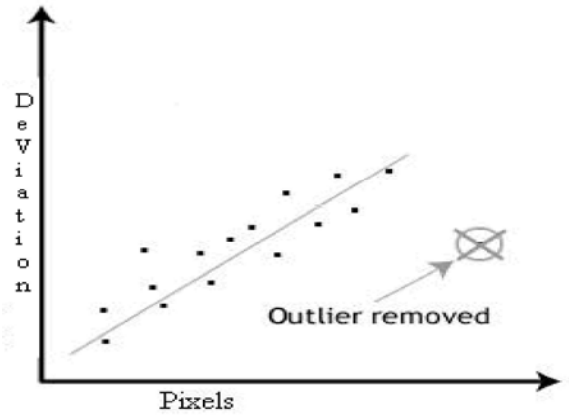

Fig. 5: Removing the detected Outlier from the image pixels

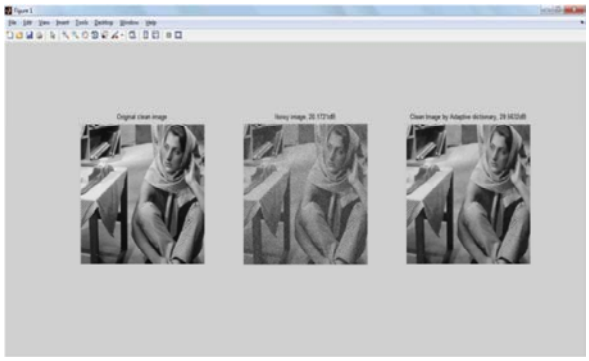

Fig. 6: Outlier detection using KSVD algorithm

Then the pixels present at the middle will be comparing with the nearby neighbour pixels values to detect the outliers presents in the image. This process continues until each and every single pixels of the image get compared. The data values will be then categories as normal data and outlier data.

Figure 5 represents the process of removing the detected outlier from the image pixels by comparing the deviations of the image pixels with nearby pixel group. Figure 6 represent detection of an outlier using KSVD algorithm. The raw black and white is taken now addition outliers will added to the image. KSVD algorithm is applied over the outlier noisy image to remove the unwanted outlier noise. KSVD algorithm shows better result to remove an outlier presented in the image when compare to the existing approaches.

CONCLUSION

The proposed method of KSVD is the new model of outlier detection in an image. It functions intuitively and predicts outlier's presents on the data objects. The pixels data objects are clustered and their likelihood values are calculated, then data objects are classified into normal and outlier group in order to detect the outlier. The proposed work focuses on Quality of the image consisting of various real time applications and the problem of imperfect data labeling on dynamic change of data set is solved by this approach. This shows that the proposed work is better than previous mechanism and tradeoffs between false detection and time duration. Future work of this model can be extended in identifying the outliers in large data set of dynamic online streaming environment.

REFERENCES

1. Grubbs, F.E., 1969. Procedures for detecting outlying observations in samples, Technometrics.
2. Chandola, V., A. Banerjee and V. Kumar, 2009. Anomaly detection: A survey, ACM CSUR., 41(3).
3. Lee, Y.J., Y.R. Yeh and Y.C.F. Wang, 2012. Anomaly detection via online over-sampling principal component analysis, IEEE Trans.Knowl, Data Eng., 25(7): 1460-1470.
4. Eskin, E., 2000. Anomaly detection over noisy data using learned probability distributions. Proc. ICML, 2000. San Francisco, CA, USA, pp: 255-262.
5. Chen, F., C.T. Lu and A.P. Boedihardjo, 2010. GLS-SOD: A generalized local statistical approach for spatial outlier detection, Proc. ACM SIGKDD Int. Conf. KDD.2010. New York, NY, USA, pp: 1069-1078.
6. Hido, S., Y. Tsuboi, H. Kashima, M. Sugiyama and T. Kanamori, 2011. Statistical outlier detection using direct density ratio estimation, Knowl. Inform. Syst., 26(2): 309-336.
7. Jiang, S.Y. and Q.B., 2008. An Clustering-based outlier detection method. Proc. ICFSKD.2008. Shandong, China, pp: 429-433.
8. Smith, R., A. Bivens, M. Embrechts, C. Palagiri and B. Szymanski, 2002. Clustering approaches for anomaly based intrusion detection, Proc. Intell. Eng. Syst. Artif. Neural Netw, pp: 579-584.
9. Shi, Y. and L. Zhang, 2011. COID: A cluster-outlier iterative detection approach to multi-dimensional data analysis, Knowl. Inform.Syst., 28(3): 709-733.
10. Ghoting, A., S. Parthasarathy and M.E.Otey, 2008 Fast mining of distance-based outliers in high-dimensional datasets, Data Min. Knowl. Discov., 16(3): 349-364.

11. Ghoting, A., S. Parthasarathy and M. Otey, 2008. Fast mining of distance-based outliers in high-dimensional datasets, Data Min.Knowl. Discov., 16(3): 349-364.
12. Angiulli, F. and F. Fassetti, 2009. Dolphin: An efficient algorithm for mining distance-based outliers in very large datasets,. ACM Trans. Knowl. Discov. Data, 3(4): 1-57.
13. Niennattrakul, V., E.J. Keogh and C.A. Ratanamahatana, 2010. Data editing techniques to allow the application of distance-based outlier detection to streams, Proc. IEEE ICDM. 2010. Sydney, NSW, USA, pp: 947-952.
14. Bhaduri, K., B.L. Matthews and C. Giannella, 2011. Algorithms for speeding up distance-based outlier detection, Proc. ACM SIGKDD Int. Conf. KDD, New York, NY, USA, pp: 859-867.
15. Breunig, M.M., H.P. Kriegel, R.T. Ng and J. Sander, 2000. LOF: Identifying density-based local outliers, Proc. ACM SIGMOD Int. Conf. Manage. Data. 2000. New York, NY, USA, pp: 93-104.
16. Tax, D.M.J. and R.P.W. Duin, 2004. Support vector data description.Mach. Learn, 54(1): 45-66.
17. Tax, D.M.J., A. Ypma and R.P.W. Duin, 1999. Support vector data description applied to machine vibration analysis, Proc. ASCI, pp: 398-405.
18. Jordaan, E.M. and G.F. Smits, 2004. Robust outlier detection using SVM regression, Proc. IJCNN, pp: 1098-1105.
19. Spiros Papadimitrious, Hiroyuki Kitagawa and B. Phillip Gibbons, 2003. LOCI: Outlier detection using the local correlation integral, International conference on data engineering.
20. Liu, Bo, Yanshan Xiao, Philip S. Yu, Zhifeng Hao and Longbing Cao, 2014. An Efficient Approach for Outlier Detection with Imperfect Data Labels, IEEE Transaction on knowledge and data engineering, 26(7).
21. Elpiniki, I., 2013. Papageorgiou Member IEEE and Dimitris, Iakovidis K, Intuitionistic Fuzzy Cognitive Maps, IEEE Transaction on Fuzzy systems, 21(2).
22. Priyam Chatterjee, 2007. Image Processing and Reconstruction, Denoising using the K-SVD Method.