CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1. CONCLUSION

Outlier detection is an important data mining task that has attracted substantial attention within diverse research communities and the areas of application. By now, many techniques have been developed to detect outliers. However, most existing research focuses on numerical data. And they cannot directly apply to categorical data because of the difficulty of defining a meaningful similarity measure for categorical data. Optimization-based model is proposed for categorical outlier detection, which a new concept of weighted holoentropy which captures the distribution and correlation information of a dataset. To solve the optimization problem they derive a new outlier factor function from the weighted holoentropy and show that computation/updating of the outlier factor can be performed without the need to estimate the joint probability distribution. The works carried out in the recent work either support unsupervised learning for categorical data and the measurement of the attributes consider for entire data without consideration of the attribute value. Another one of the main problem of outlier detection algorithms are data sets with non-homogeneous densities.

To solve this problem in the initial stage of the work, hybrid Expectation Maximization methods which combines the procedure of the Ascent Monte Carlo method so it is named as Ascent-Based Monte Carlo Expectation– Maximization (AMCEM) is proposed for outlier detection. The proposed weighted Kullback Leiber Divergence(KLD) measure the attribute value with maximum likelihood of outlier candidates, while the efficiency of the algorithms results from the outlier factor function. The outlier factor of an object is solely determined by the object and its updating does not require estimating the data distribution. The proposed AMCEM also estimate an ascent property for the number of outliers and an anomaly candidate set. This bound, obtained under a very reasonable hypothesis on the number of possible outliers, allows us to further reduce the search cost.
Here the outlier detection is now formulated be stated as an optimization problem. Moreover, one might how to determine the optimal value of outliers in the original dataset becomes difficult task. A possible theoretical approach to this problem is to search for a range of values of outliers. For this purpose second stage of the work Particle Swarm Optimization (PSO) is introduced to the optimization problem.

In the second stage of the work, Expectation Maximization with Particle Swarm Optimization Based Weighted Clustering (EMPWC) is proposed for outlier detection. Here the efficiency of EMPWC outlier detection algorithm majorly depends on attribute frequency which is derived from weighted entropy optimization .The weighted entropy optimization considers both the data Shannon and Jensen-Shannon Divergence (JSD) to measure the likelihood of outlier candidates. At first algorithm appears to be a straightforward generalization of standard EM for Gaussian mixtures, the second one is performed based on weight computation results. In this model weight values are derived from entropy measures. In addition this work, range values of outliers (o) are optimized using PSO. Also estimate an upper bound for the number of outliers and an anomaly candidate set. This bound, obtained under a very reasonable hypothesis on the number of possible outliers, allows us to further reduce the search cost. Based on this PSO method the data clustering results are increased and hence the algorithm is extremely efficient.

The major disadvantages of PSO algorithm are that it is easy to fall into local optimum in high-dimensional space and has a low convergence rate in the iterative process. To solve this problem Bat Algorithm is introduced in this work. However missing values is also another common problem to every real world data. Data Normalization standardize the raw data by converting them into specific range using linear transformation which can generate good quality clusters and improve the accuracy of clustering algorithms.

In the final stage of the work, new Graph Based Semi-Supervised Clustering with Bat Algorithm (GSSBAT) is proposed for outlier detection. The major important steps in the final stage of the work are preprocessing, outlier detection and clustering. The preprocessing is done with the help of min-max normalization
approach. It is more useful for dealing with the missing values and hence the clustering performance is progressed than the previous method. Then the unbalanced dataset problem is handled by using SMOTE with kNN approach which is focused to increase the dataset efficiency. Apply the BAT optimization algorithm for optimizing the outlier attributes. By computing the optimal fitness function, the outlier attributes are identified more effectively.

In this research work, the range values of outliers (o) get optimized employing the BAT algorithm. On the basis of this Bat technique, the data clustering results are seen to increase and therefore the algorithm proves to be greatly efficient. Then use the GSS based clustering algorithm for accurate clustering. The proposed clustering algorithms are specifically applied for University of California, Irvine (UCI) machine learning repository. The result proves that the proposed GSSBAT approach has superior performance in terms execution, Normalized Mean Square Error (NMSE), Detection Rate (DR) and False Alarm Rate (FAR) than the previous approaches. From the results it concludes that the proposed GSSBAT algorithm has lesser average FAR value of 15% which is 2.33%, 3.75%, 8.066%, and 10.21% lesser value when compared to EMPWC, AMCEM, ITB-SS, and ITB-SP methods respectively.

6.2. FUTURE WORK

Future research includes further improving the speed and extending for distributed datasets. In the present work, the datasets on which the proposed technique is tested are of integer or real type. So, our work is undergoing to extend the current work of the algorithm to work other type datasets.

On a revolutionary perspective, assessments is carried out on a small actual data set and a bundle of artificial data sets indicate that the algorithms do proposed attempt to have the optimization of the selection of candidates to be outliers. In addition, the experiments over real and artificial data sets compared with other algorithms assure the efficiency and purposefulness of the proposed algorithms that are in practice. In future, the kernel based outlier detection method can be developed for distributed mixed arbitrary-type data sets.