DOI:10.26524/krj135

Kong. Res. J. 3(1) : 81-83, 2016 ISSN 2349-2694
Kongunadu Arts and Science College, Coimbatore.

AN EFFICIENT FUZZY BASED ANOMALY DETECTION USING COLLECTIVE CLUSTERING ALGORITHAM

Gomathi. K[1*] and R.Umagandhi[2]
[1]Department of Computer Science, [2]Department of Computer Technology, Kongunadu Arts and Science College, Coimbatore - 641 029.
*E.mail: gomathi.karupuswamy@gmail.com

ABSTRACT

Anomaly detection is a significant problem that has been researched within various research areas and application domains. Many anomaly detection methods have been particularly examined for certain application domains, as others are more standard. This present study describes an anomaly detection technique for unsupervised data sets accurately reduce the data from a kernel Eigen space performing a batch re-computation. For each anomaly behavior activities is to identify the key factors, which are used by the methods to differentiate between normal and abnormal actions. This present study provides a best and brief understanding of the techniques belonging to each anomaly and kernel mapping category. Further, for each grouping, to identify the improvements and drawbacks of the techniques in that category. It also provides a discussion on the computational complexity of the techniques since it is an important issue in real application domains hope that this survey will provide a good understanding of the many directions in which research has been done on this topic.

Keywords: Adaptive, non-stationary, anomaly detection, outlier detection, kernel principal component analysis, kernel methods.

1. INTRODUCTION TO DATA MINING

Data Mining, "The Extraction of hidden predictive information from large databases", is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. The databases for hidden patterns are finding predictive information that experts may miss because it lies outside their expectations.

Data Mining consists of more no of collecting and managing data; it also includes analysis and prediction. Data Mining can be performed on data represented in quantitative, textual, or multimedia forms. Data Mining applications can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering, and forecasting. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis.

Data Mining techniques are the result of a long process of research and product development(Kriegel, Et al,2008). This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

2. MATERIALS AND METHODS

The proposed architecture accepts the user parameters as input which contains the MATLAB simulation where the Fuzzy based kernel mappings with adaptive Neighboring Splitting and Merging algorithm is applied to the datasets. This architecture in figure 3.1 follows a path from the start to end state. The users initialize the number of *k*-value as cluster parameters in which the anomaly detection process is to be evaluated(Zhang,et al.,2010).

Anomaly detection holds great potential for detecting previously unknown outliers (Barnett and Lewis, 1994). In order to be effective in a practical environment, anomaly detection systems have to be capable of online learning and handling concept. Clustering becomes difficult due to the increasing sparsity of such data, as well as the increasing difficulty in distinguishing distances between data points. It has been widely recognized that consensus clustering can help to generate robust clustering results, find bizarre clusters, handle noise, outliers and sample variations, and integrate solutions from multiple distributed sources of data or attributes. An presents an optimal perspective on the problem of kernel principal component analysis in high-dimensional data (Ding and Kolaczyk,2013). The proposed method called "Fuzzy based kernel mappings with adaptive Neighbouring Splitting and Merging" (FKANSM), which takes as key measures of correspondence between pairs of data points. The proposed method is to establish a unified framework for FKANSM on both supervised and unsupervised data sets. Also, we examine some important factors, such as the clustering quality and assortment of basic partitioning, which may affect the performances of FKANSM. Experimental results on various synthetic and real world data sets demonstrate that FKCNC is highly efficient and is equivalent to the state-of-the-art methods in terms of clustering index quality. In addition, FKANSM shows high robustness to incomplete basic partitioning with many anomaly values.

3. RESULTS

To find the clusters of a data set sampled from a certain unknown distribution is important in many machine learning and data mining applications. Probability density estimate may represent the distribution of data in a given problem and then the modes may be taken as the representatives of clusters. As a nonparametric method, the kernel density estimation is mostly applied in practice to model the unknown probability density function (Fig. 1).

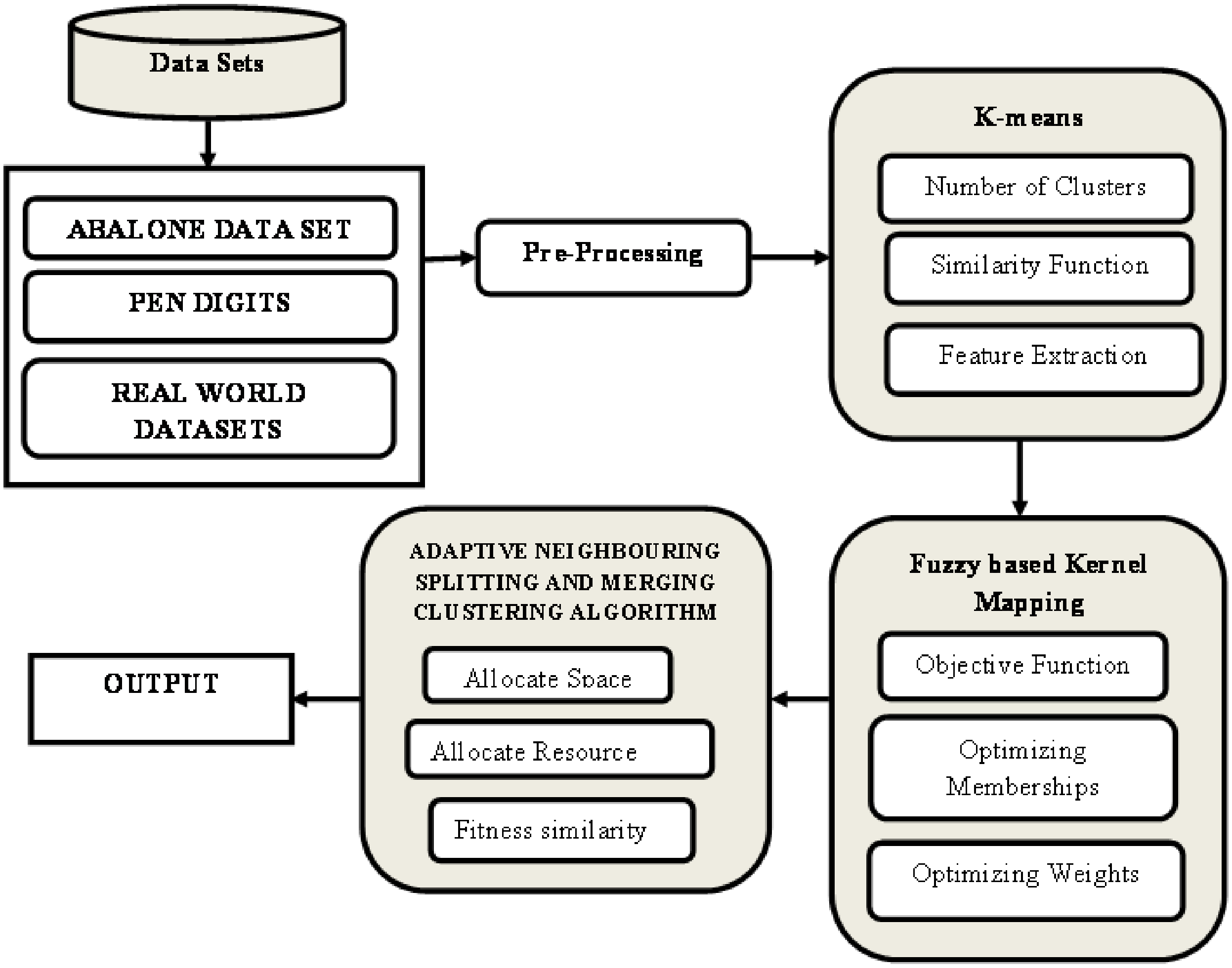
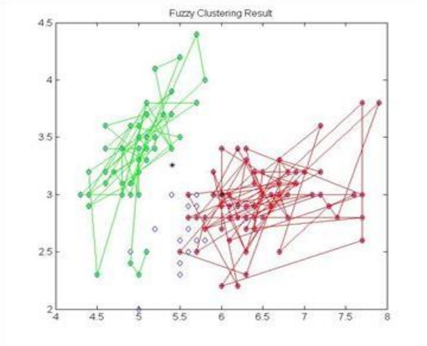


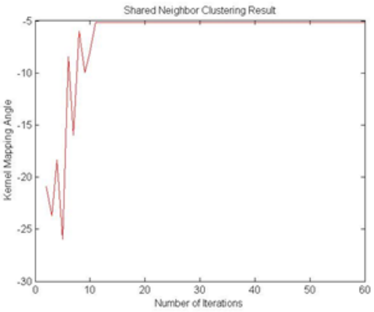
Fig. 1. Architecture of Proposed System

3.1. Abalone Dataset

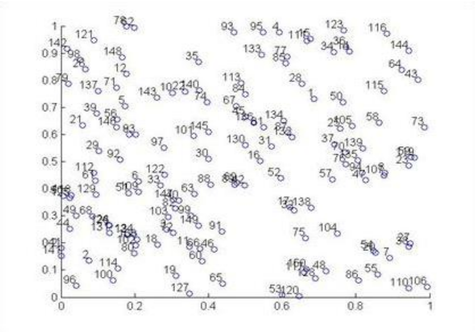
The abalone dataset describes the predicting age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope - a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem. From the original data examples with missing values were removed (the majority having the predicted value), and the ranges of the continuous values have been scaled for use with an "Artificial Neural Network" (ANN) (by dividing by 200).



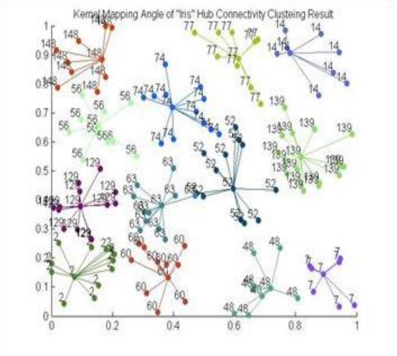
(a) Outliers (unwanted data)



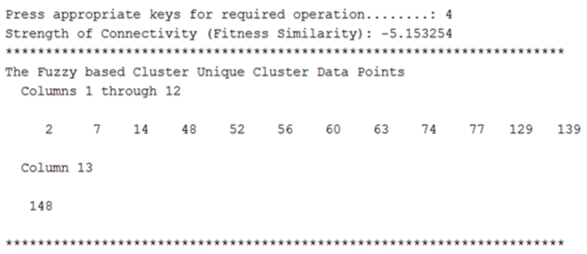
(b) Outlier range



(c) Data sets



(d) Cluster data



```
Press appropriate keys for required operation........: 4
Strength of Connectivity (Fitness Similarity): -5.153254
**********************************************************************
The Fuzzy based Cluster Unique Cluster Data Points
  Columns 1 through 12

     2     7    14    48    52    56    60    63    74    77   129   139

  Column 13

   148

**********************************************************************
```

(e) Fitness similarity for data points

Fig. 2 (a-e). Abalone datasets

4. CONCLUSION

Anomaly detection in data mining area is efficient and effective task to ensure the quality and right decisions. A selection of anomaly detection models was proposed in an Efficient Fuzzy Based Anomaly Detection; however, most of them suffer from high dimensional datasets effectiveness or high outliers. This result shows the challenges that face the design of an efficient and effective anomaly detection model for synthetic and real-world data sets in data mining domain that should be satisfied to design such models.

REFERENCES

Kriegel, H.P., A. Zimek and M. Schubert, (2008). *Angle-based outlier detection in high-dimensional data. In*: Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 444–452.

Zhang, Y., N. Meratnia and P.J. Havinga, (2010) Ensuring high sensor data quality through use of online outlier detection techniques, *Int. J. Sens. Netw*. **7**(3): 141–151.

Barnett, V. and T. Lewis, (1994). *Outliers in Statistical Data*, New York, USA: Wiley (3).

Ding, Q. and E.D. Kolaczyk, (2013) A compressed PCA subspace method for anomaly detection in high-dimensional data. *IEEETrans. Inf. Theory* **59**(11):7419–7433.