ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

An Efficient Anomaly Detection using Fuzzy based Adaptive Neighbouring Splitting and Merging Clustering

V. Kalai Selvi[1], M. Sheela Newsheeba[2]

M.Phil Research Scholar, Nehru Arts and Science College, Thirumalayampalayam, Coimbatore, India

Assistant Professor, Nehru Arts and Science College, Thirumalayampalayam, Coimbatore, India

ABSTRACT: Unsupervised data clustering arises obviously in a lot of applications, and have regularly presented a great covenant with for usual data mining techniques. In this paper, presents an optimal view on the problem of anomaly detection in high-dimensional data. The proposed method called "*Fuzzy based kernel mappings with Adaptive Neighboring Splitting and Merging (FKANSM)*", which takes as key measures of correspondence between data elements. The proposed method is to establish a unified framework for *FKANSM* on both unsupervised and semi-supervised data sets. Meanwhile, the proposed system examine some important factors, such as the clustering quality and variety of basic partitioning, which may affect the performances of *FKANSM*. Experimental results on various synthetic and real world data sets demonstrate that *FKANSM* 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.

KEYWORDS: Fuzzy logic, anomaly detection, High dimensionality, kernel mapping

I. INTRODUCTION

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 [1]. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data Mining consists of more than 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 [2]. 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. 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.

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters [7]. Data modelling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept.

Therefore, clustering is unsupervised learning of a hidden data concept and the fuzzy clustering is the most widely used technique for hidden data analysis.

II. Related work

In [3] authors proposed a new method for performing a nonlinear form of Principal Component Analysis is proposed. By the use of integral operator kernel functions, one can efficiently compute principal components in high-dimensional feature spaces, related to input space by some nonlinear map; for instance the space of all possible 5-pixel products in images. To give the derivation of the method and present first experimental results on polynomial feature extraction for pattern recognition. In [4] provide algorithms for adding and subtracting eigenspaces, thus allowing for incremental updating and down dating of data models. Importantly, and unlike previous work, we keep an accurate track of the mean of the data, which allows our methods to be used in classification applications. The result of adding eigenspaces, each made from a set of data, is an approximation to that which would obtain were the sets of data taken together. Subtracting eigenspaces yields a result approximating that which would obtain were a subset of data used. In [5] authors proposed a classifier ensemble using Winnow. For the constant-update strategy we used the nearest neighbor with a fixed size window and two methods with a learning rate: the online perception and an online version of the linear discriminate classifier (LDC). For the detect-and-retrain strategy we used the nearest neighbor classifier and the online LDC. In [6] authors proposed outlier detection techniques based on one-class quarter-sphere support vector machine meeting constraints and requirements of WSNs. To reduce the false alarm rate while increasing the detection rate and to enable collaborative outliers detection, we take advantage of spatial and temporal correlations that exist between sensor data.

III. Proposed algorithm

A. *FEATURE SELECTION*

A “feature selection” refers to a portion of the data points. Typically before collecting data, features are specified or preferred. Features can be discrete, continuous, or insignificant. Feature selection for high-dimensional data clustering is the task of disregarding irrelevant and redundant terms in the vectors that represent the hubs, aiming to find the smallest subset of terms that reveals “natural” clusters of hubs. To Searching for the small subset of relevant terms will speed up the clustering process, while avoiding the curse of dimensionality.

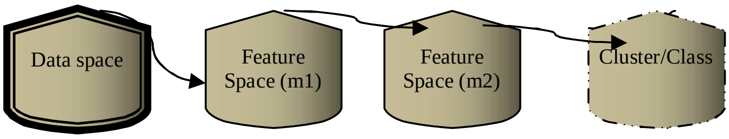

Fig: 1. Feature Selection

The Noise filter removes irrelevant features using a modified form of the Relief algorithm, which assigns relevance values to features by treating training samples as points in feature space. For each sample, it finds the nearest “hit” (another sample of the same class) and “miss” (a sample of a different class), and adjusts the significance value of each feature according to the square of the feature difference between the sample and the hit and miss. Noise Filter feature selection methods evaluate attributes prior to the learning process, and without specific reference to the clustering algorithm that will be used to generate the final result. The filtered dataset may then be used by any clustering algorithms.

B. *ANOMALY DETECTION KERNEL MAPPING DATA CLUSTERS*

A correlation between low data elements (i.e., low feature) and anomalies was also observed. A low-data points score indicates that a point is on average far from the rest of the points and hence probably an outlier. In high-dimensional spaces, however, low data point elements are expected to occur by the very nature of these spaces and data resource. The kernel mapping can be applied using more general notions of similarity, and the similarities may be positive or negative. The output of the algorithm is unchanged if the similarities are scaled and/or offset by a constant (as long as the preferences are scaled and/or offset by the same constant). To compute fitness measure over the set of possible clusters and then chooses among the set of cluster candidates points (hubs) those that optimize the measure used. To identify the cluster of a specific vertex or to group all of the vertices into a set of clusters, and then present possible cluster fitness measures that serve for methods that produce the clustering by comparing different groupings and selecting one that meets or optimizes a certain criterion. The ratio of the cluster is to minimum sums of degrees either inside the cluster or outside it. A fitness function is evaluated for all neighbours and the outcome is used to choose to which neighbour the search will proceed.

C. *KERNEL MAPPING ESTABLISHMENT*

The degree of branching can be specified with a kernel k that is directly applied to the similarity matrix. It is shown that the generated clusters can still be monotonic depending on the used linkage measure even though the induced dissimilarity measures are no longer ultra metrics. Using the pair-wise merged clusters; an additional shrinking process is proposed to generate topic related groups with more than two cluster elements.

- The process of determining the degree to which a value belongs in a kernel set
- The value returned by a shared-Neighbour cluster
- Most variables in a hub-based system have multiple data points attached to them
- Kernel mapping that variable involves passing the crisp value through each neighbour attached to that value

Here unsupervised dataset is an object matrix. Clusters are groups of similar data elements. Resemblance coefficient represents the degree of similarity and non similarity between the items. The main aim of clustering analysis is identify and quantification of these architecture elements. Identifying the membership and location center of the clusters is main process in the cluster analysis. Some time data in the cluster is well packed. But due to the complex nature of the components the data may not be packed well in the clusters. Some of the elements lie outside the cluster region.

D. *ANOMALY-FUZZY COLLECTIVE CLUSTERING ALGORITHM*

The collective clustering algorithm works message passing among data points. Each data points receive the availability from others data points (from pattern) and send the responsibility message to others data points (to pattern). Sum of responsibilities and availabilities for data points identify the cluster patterns.

The high-dimensional data point availabilities A(i, k) are zero: A(i, k) = 0, R(i, k) is set to the input similarity between point i and point k as its pattern, minus the largest of the similarities between point i and other candidate patterns. Fuzzy approach computes two kinds of messages exchanged between data points. The first one is called "responsibility" $r(i, j)$: it is sent from data point i to candidate exemplar point j and it reflects the accumulated evidence for how well-suited point j is to serve as the exemplar for point i. The second message is called "availability" $a(i, j)$: it is sent from candidate exemplar point j to point i and it reflects the accumulated evidence for how appropriate it would be for point i to choose point j as its exemplar. At the beginning, the availabilities are initialized to zero: $a\ (i, j) = 0$. The update equations for $r\ (i, j)$ and $a(i, j)$ are written as,

$$r(i,j) = s(i,j) - \max_{j' \neq j}\{a(i,j') + s(i,j')\} \qquad \text{eqn. (1)}$$

$$a(i,j) = \begin{cases} \min\{0, r(j,j) + \sum_{i' \neq i,j} \max\{0, r(i',j)\} & , i \neq j \\ \sum_{i' \neq i} \max\{0, r(i',j)\}, & i = j \end{cases} \qquad \text{eqn. (2)}$$

IV. SIMULATION RESULTS

The Proposed work has been evaluating the performance of the incremental anomaly detection technique, real-world data sets are used. Evaluation of anomaly detection techniques often uses two (or more) class data sets in the evaluation of performance. One (or more) of the classes is used as the normal class, and the remaining classes are used as the anomalous data [8]. The performance on the MNIST data set is slightly lower than that of the dynamic sliding window approach. However, it is able to significantly reduce the number of updates that are required. The false positive rate (FPR) is computed as the ratio of false positives to normal measurements and the true positive rate (TPR) is the ratio of true positives to anomalous measurements. To compare schemes, receiver operating characteristic (ROC) curves are generated by varying the anomaly ratio used to determine the threshold distance for the reconstruction error. The threshold, the fraction of the training data rejected, was varied from 0.01 to 0.99 in steps of 0.01. The resulting FPR and TPR form the ROC curve. Results were compared for various predefined numbers of clusters in algorithm calls. Each algorithm was tested 50 times for each number of clusters. Neighborhood size was 2 to 16.

Table1: Clustering Quality on the UCI machine learning datasets

K	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**
U_c	0.0556	0.506	0.111	0.1212	0.1488	0.3767	0.3122	0.0352
U_H	0.4296	0.0661	0.1476	0.4628	0.4039	0.5702	0.4743	0.4296
U_{cos}	0.111	0.4359	0.7352	0.5814	0.5322	0.0421	0.1448	0.3647
NU_H	0.5470	0.7069	0.0537	0.1336	0.4938	0.3619	0.4093	0.2412
FKANSM	0.5582	0.6894	0.5992	0.5863	0.6321	0.5769	0.4956	0.4723

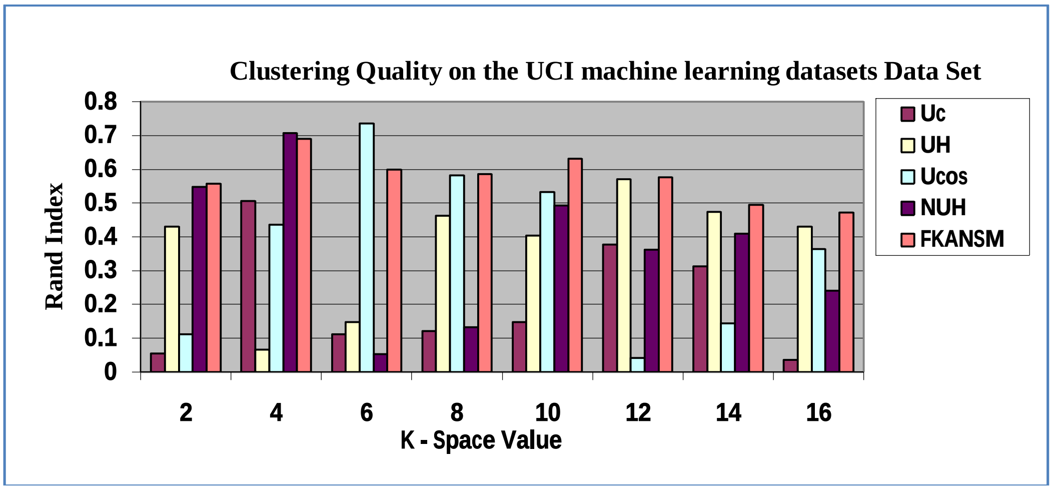

Fig.2. Clustering Quality Comparison Chart

V. Conclusion and Future Work

In this paper presents fuzzy based kernel mapping to approximate local data centers is not only a feasible option, but also frequently leads to improvement over the centroid-based approach. The proposed the Fuzzy based kernel mappings with adaptive Neighboring Splitting and Merging (FKANSM) algorithm for the Anomaly Detection is in core variations of fuzzy based clustering algorithm using different weight measures applied to the vector of base-level clustering's baseline on both synthetic and real-world data, as well as in the presence of high levels of artificially introduced noise. This initial evaluation suggests that using data points both as cluster prototypes and points guiding the centroid-based search is a promising new idea in clustering high-dimensional and noisy data. Also, global data point estimates are generally to be preferred with respect to the local ones. To apply to the real data sets we need to refine the adjacency matrix by the hard-thresholding, say, and this area is worth pursuing as future research.

References

1. I. Jolliffe, Principal Component Analysis. Hoboken, NJ, USA: Wiley, 2005.
2. Q. Ding and E. D. Kolaczyk, “A compressed PCA subspace method for anomaly detection in high-dimensional data,” IEEE Trans. Inf. Theory, vol. 59, no. 11, pp. 7419–7433, Nov. 2013.
3. B. Scholkopf, A. Smola, and K.-R. M€uller, “Nonlinear component analysis as a kernel eigenvalue problem,” Neural Comput., vol. 10, no. 5, pp. 1299–1319, 1998.

4. P. Hall, D. Marshall, and R. Martin, “Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition,” Image Vis. Comput., vol. 20, no. 13, pp. 1009–1016, 2002.
5. J. J. Rodr ıguez and L. I. Kuncheva, “Combining online classification approaches for changing environments,” in Proc. Int.Workshop Struct., Syntactic, Statist. Pattern Recognit, 2008, pp. 520–529.
6. Y. Zhang, N. Meratnia, and P. J. Havinga, “Ensuring high sensor data quality through use of online outlier detection techniques,” Int. J. Sens. Netw., vol. 7, no. 3, pp. 141–151, 2010.
7. I. B. Khediri, M. Limam, and C. Weihs, “Variable window adaptive kernel principal component analysis for nonlinear nonstationary process monitoring,” Comput. Ind. Eng., vol. 61, no. 3, pp. 437–446, 2011.
8. Y. Lee, Y. Yeh, and Y. Wang, “Anomaly detection via online oversampling principal component analysis,” IEEE Trans. Knowl. DataEng., vol. 25, no. 7, pp. 1460–1470, Jul. 2013.