ISSN 2047-3338

Classification of Dataset Using Clustering Technique

Satyanarayan Misra[1], Prof. Sanjay Singh[2] and Pradeep Kumar Tiwari[3]
[1]JJT University Jaipur, India
[2]AMITY University Campus, Lucknow, India
[3]DAMS Kanpur, India
[1]misra.satyanarayan@gmail.com, [2]sanjaysingh10@rediffmail.com, [3]tprakhar.tiwari18@gmail.com

***Abstract*– In this paper we establish a different implementation level clustering technique which provides a new dimension for classification of datasets. The training set and the testing set of data are classified according to the separate kind of algorithms discussed here. The analysis of the performance will show a clear edge over the other existing technique used for data classification. The future research issues which need to be resolved and investigated further are given with new trends and ideas.**

***Index Terms*– Pattern Recognition, Feature Extraction, Neighborhood Technique, Data Clustering and Data Classification**

I. INTRODUCTION

PATTERN recognition can also be seen as classification process. Its ultimate goal is to optimally extract patterns based on certain conditions and is to separate one class from the others. Pattern recognition was often achieved using linear and quadratic discriminants [1] the k-nearest neighbor classifier [2], or the Parzen density estimator [3], template matching [4] and neural networks [5]. These methods are basically static. The problem of using these recognition methods have to construct a classification rule without having any idea of distribution of the measurements in different group. The randomized hybrid technique is a concept in the implementation level which combines the idea of quadratic discriminants and the k-nearest neighbor classifier with a little modification.

This paper is organized as follows: we first introduce some of the general process of pattern recognition in section 2. Next section will be a brief discussion on the motivation and overview. The implementation level algorithm and performance evaluation of the proposed algorithm will be discussed in section 4 and 5. Conclusions and future proceedings will be followed in subsequent sections.

II. GENERAL PROCESS OF PATTERN RECOGNITION

A Pattern is a pair comprising an observation and a meaning. Pattern Recognition is inferring meaning from observation. Designing a pattern recognition system is establishing a mapping from measurement space into space of potential meanings, whereby the different meanings are represented in this space as discrete target points. The basic components in pattern recognition are preprocessing, feature extraction and selection, classifier design and optimization.

A. Preprocessing

The role of preprocessing is to segment the interesting pattern from the background. Generally noise filtering, smoothing and normalization should be done in this step. The preprocessing also defines a compact representation of the pattern.

B. Feature Selection and Extraction

Feature should be easily computed, robust, insensitive to various distortions and variations in the images and rotationally invariant. Two kinds of features are used in pattern recognition problems [7]. One kind of features has clear physical meaning, such as geometric or structural and statistical features. Another kind of features has no physical meaning. We call these features as mapping features. The advantage of physical features is that they need not deal with irrelevant features.

The advantage of mapping features is that they make classification easier because clear boundaries will be obtained between classes but increasing the computational complexity. Feature selection is to select the best subset from the input space. Its ultimate goal is to select the optimal feature subset that can achieve the highest accuracy results. While feature extraction is applied in the situation when no physical features can be obtained [13].

In feature extraction, most methods are supervised. These approaches need some prior knowledge and labeled training samples. There are two kinds of supervised methods used: Linear feature extraction and Nonlinear feature extraction. Extraction techniques like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) etc are listed under Linear and PCA Network; Multidimensional Scaling (MDS) are listed under Nonlinear Extraction.

C. Classifier Design

After optimal feature subset is selected a classifier can be designed using various approaches. There are various existing techniques [6] like based on similarity approach, probabilistic approach, and decision boundaries approach and clustering approach etc. Decision tree algorithms are a very popular technique for learning patterns from data and using these patterns for classification. In [8], we report the application of decision tree algorithms to computer intrusion detection. Some of the existing decision tree algorithms support the incremental learning [11]-[9]. Some other data mining algorithms support the scalability [12]-[10]. However, in this paper we will concentrate on the mixed approach on similarity and clustering technique which will provide a excellent classifier to distinguish the different data sets available in the test sets. The classifier will be first obtained with respect to the training sets data and will give an optimized boundary for the test sets.

D. Optimization

The optimization is not a separate step, it is combined with several parts of the pattern recognition process. In preprocessing, optimization guarantee that the input pattern have the best quality. Then in feature selection and extraction part, optimal feature subsets are obtained under some optimization techniques. Furthermore the final classification error rate is lowered in the classification part.

III. DESIGNING OVERVIEW

Let us consider a set of points which are regarded as training sets. We first select randomly a data point and also choose a neighborhood radius. The distance of other data points to the randomly selected data point is calculated. If the calculated distance is less than the pre-specified neighborhood radius then the data point is included in the cluster region of the randomly selected point, else it is classified in another cluster. This process will be continued till all the data elements in the training set are not completed. Again the whole process can be again repeated to get another randomly selected point as cluster centre. This is well explained in the Fig. 1.

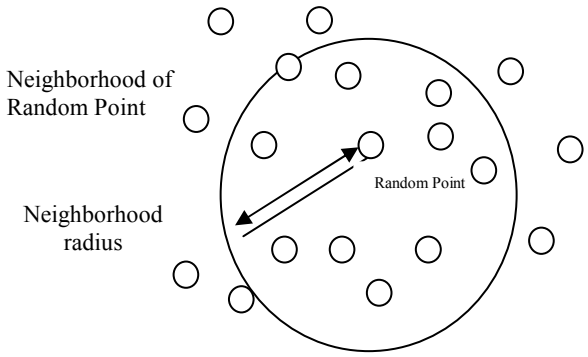

Fig. 1: Classification of Training Sets

The classification rule depends on the neighborhood radius and the number of selection of random data points. Once the cluster centre is chosen as appropriately, the testing sets will be classified accordingly.

IV. ALGORITHM

Let T= $\{a_i|\ 1\le i\le n\}$ be the data elements of the training set and ε be the distance measure for the neighborhood approach.

- *Learning on Training Set For 2 characteristics per data point i.e., x and y*

1. Choose randomly a data point from the training set say a_r and include in the cluster region R_r i.e. $R_r= \{a_r\}$.
2. for all other data point $a_j\in$ T, $1\le j\le n$ and $j\ne r$

2.1. Apply

$$d(a_j,a_r)=\left[[a_j(x)-a_r(x)]^2+[a_j(y)-a_r(y)]^2\right]^{1/2}$$

2.2. If $d(a_j,a_r)\le\varepsilon$ then

$R_{ir}=R_{ir}\cup a_j$ else $R'_{ir}=R'_{ir}\cup a_j$.

3. Repeat step 1 and 2 for k-times, $1\le k\le n$,where $n=|T|$.
4. Find $\max\left(|R_{ir}|\right),1\le i\le n$ and select the a_i as the learned cluster centre.

The next section follows a generalized algorithm to perform the supervised learning with more number of characteristics per data elements.

For p characteristics per data point

1. Choose randomly a data point from the training set say a_r and include in the cluster region R_r i.e. $R_r= \{a_r\}$.
2. for all other data point $a_j\in$ T, $1\le j\le n$ and $j\ne r$

2.1. Calculate

$$d_i(a_j,a_r)=\left[[a_j(x_i)-a_r(x_i)]^2\right]^{1/2}$$

$$d(a_j,a_r)=\left[\sum_{i=1}^{p}[a_j(x_i)-a_r(x_i)]^2\right]^{1/2}$$

2.2. If $d_i(a_j,a_r)\le\varepsilon_i$ and

$d(a_j,a_r)\le\varepsilon$

then $R_{ir}=R_{ir}\cup a_j$ else $R'_{ir}=R'_{ir}\cup a_j$

3. Repeat step 1 and 2 for k-times, $1\le k\le n$, where $n=|T|$
4. Find $\max\left(|R_{ir}|\right),1\le i\le n$ and select the a_i as the learned cluster centre.

This algorithm will work for selecting as best as possible cluster centre with the provided training set data. In the next algorithm 4.2 the test data will be classified.

- *Classifying Test Data Set*

Let T_e= {b_j|1≤j≤m} be the testing data set. Let X_i be the i^{th} attribute of a variable of a data point X, L_i be the i^{th} coordinate of centroid of a cluster L and C_i is the correlation coefficient of the i^{th} attribute variable and the target variable.

for all j

Calculate $d(X,L) = \sum_{i=1}^{m} \frac{|X_i - L_i|}{X_i + L_i} C_i^2$

if $d(X,L) \le \varepsilon_t$ then

$R_{ir} = R_{ir} \cup b_j$ else $R_{ir}^{'} = R_{ir}^{'} \cup b_j$.

Where ε_t be the threshold distance for decision boundaries and *d(X, L)* is the Canberra distance metric.

V. PERFORMANCE STUDY

The Comparative model is based on the statistically collected data over a particular pattern. Fig. 2 depicts a clear focus on the number of Data elements Vs Radius of Neighborhood.

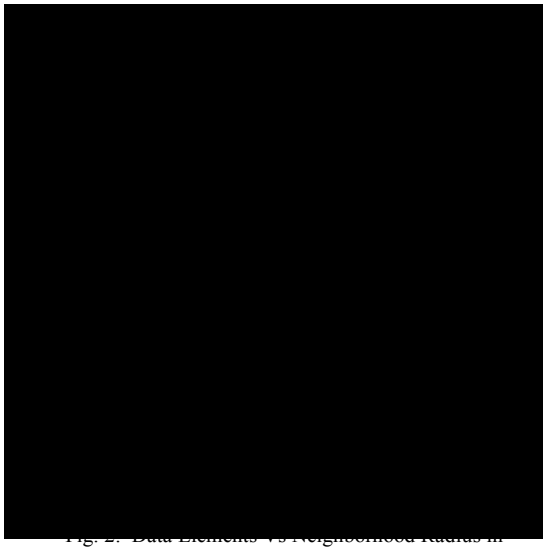

Fig. 2: Data Elements Vs Neighborhood Radius in

The graphical picture is providing a very simple inference that the pattern boundary if increases then more number of data elements will be included in the desired region. On the other hand if the neighborhood radius is decreased then the number of inclusion of data elements inside the pattern boundary is less.

The Fig. 2 is providing clear information about a comparison between the k-neighborhood technique Vs Randomized Hybrid techniques. Randomized hybrid technique is performing in a better rate as compared to the k-neighborhood classification after a fixed neighborhood range.

Fig. 3: A Comparative Study: Randomized Hybrid vs. k-Neighborhood

VI. CONCLUSION AND FUTURE WORK

The performance of Random Hybrid Technique is providing a good performance evaluation over the existing classification technique. About 3.4 - 4.7% of more number of data elements are included in the classified regions which are in the same pattern in case of Random Hybrid technique. We will extend our work to classify the pattern in a probabilistic domain. The concept of randomized Hybrid technique will be used to classify and to determine more accurately the data elements corresponds to a particular pattern in the semi and complete probabilistic domain. Fuzzy logic will be countered for determination and approximation of the provided concept for classification and optimization.

REFERENCES

[1] R.A.Fisher: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, Vol. 7, part II (1936)179-188.

[2] Dasarathy, B.V. :Minimal Consistent Set(MCS) identification for optimal nearest neighbour decision systems design.IEEE Transactions on Systems, Man and Cybernetics, Vol. 24, Issue 3, (1994), 511-517.

[3] Girolami, M., Chao He: Probability Density Estimation from Optimally Condensed Data Samples. Pattern Analysis and Machine Intelligence, IEEE Transitions, Vol. 25, Issue 10, (2003), 1253-1264.

[4] Meijer, B. R., Rules and Algorithms for the Design of templates for template matching, Pattern Recognition, Vol. 1. Conference A: Computer Vision and Applications, 11th IAPR International Conference, (1992), 760-763.

[5] Hush, D.R., Horne, B. G., "Progress in Supervised Neural Networks, Signal Processing Magazine, IEEE, Vol. 10, Issue 1, (1993), 8-39.

[6] V. Vapnik, The Nature of Statistical Learning Theory. Springer (1995).

[7] Julia Neumann, Cristoph Schnorr, SVM-based Feature Selection by Direct Objective Minimization, (2004).

[8] C. Sinclair, L. Pierce, S. Matzner, An Application of Machine Learning to Network Intrusion Detection, Proceedings of 15th Annual Computer Security Application Conference (ACSAC'99), (1999), 371-377.

[9] S.L. Crawford, Extension to the CART Algorithm, International Journal of Man-Machine Studies (1989),197-217
[10] J. Shafer, R. Agrawal, M. Mehta, SPRINT, A Scalable Parallel Classifier for Datamininig, Proceedings of the 22nd VLDB Conference, Mumbai, India (1996).
[11] P.E. Utgoff, N.C. Berkman, J.A. Clouse, Decession Tree Induction based on Efficient Tree Restructuring, Machine Learning Journal 10, (1997), 5-44.
[12] S. Goil, A. Choudhary, A Parallel Scalable Infrastructure for OLAP and Data Mining. International Symposium Proceedings on Database Engineering and Applications, IEEE, (1999), 178-186.
[13] Lihong Zheng and Xiangjian H., Classification Techniques in Pattern Recognition, Conference proceedings ISBN 80-903100-8-7,WSCG'2005, January 31-February 4, 2005.

Satyanarayan Misra research scholar in JJT University, Jaipur, India. His research area is networking and testing of software.
Email: misra.satyanarayan@gmail.com

Prof. Sanjay Singh working as a professor in Amity Institute of Information Technology, Amity University, Lucknow Campus, Lucknow, he has author of many books related to computer science and published many International and national papers in various reputed journals.
Email: sanjaysingh10@rediffmail.com

Pradeep Kumar Tiwari working as a Lecturer in Dayanand Academy of Management Studies Kanpur, his research area is Parametric Programming and Networking.
Email: tprakhar.tiwari18@gmail.com