MRF: Multivariate Data Clustering using Heuristic Data Intensive Computing and Relevance Feedback Learning Approach

M. Sankara Prasanna Kumar, A. P. Siva Kumar, K. Prasanna

ABSTRACT--- Most of the problems in the real world are multivariate i.e., involves many variables. Multivariate data comprises of several datasets with more than one variable. Multivariate datasets has power to change the use of data dramatically as database size increases and it shows adequate results on predicting the effect on change in one variable will have on other variable. These datasets consist of transitive and intrinsic hidden relationships among the variables such as analyzing a variable is influenced by other process variables and preferences. It is the situation where efficient multivariate data analysis techniques exhaustively needed to catalog the given type of data. In the literature several techniques are proposed and analyzed; one such technique is multivariate data clustering.

This paper will present a unified framework of multivariate data clustering using heuristic data intensive computing and relevance feedback learning. The implementation starts with formalizing a heuristic data intensive computing (HDIC) which have the ability to handle data flows. Clustering the data is performed with proposed Relevance feedback learning algorithm with consensus functions. These functions are selected as the change in the cluster ensemble selection, combine and reduction. In this proposed approach we have used a new kind of distance functions such as Camberra, Chi-square and Cityblock. The empirical analysis shows that, the proposed approach attains better cluster ensembles on various multivariate datasets taken from UCI and out performs with k-nearest neighbour (KNN) in different settings. The performance of the proposed approach is assessed with Accuracy and F1-measure.

Index Terms—Multivariate data, clustering, consensus functions, cluster ensembles, k-nearest neighbour (KNN).

I. INTRODUCTION

Since from last two decades, the developments in information technology has introduced the necessity of data processing techniques. Similarly the developments in database technology, has introduced various kinds of databases and forcing us to develop new techniques and analysis methods for efficient data prediction and decision making.

Multivariate data or "data with more than one variable" is an active research area with a wide range of applications [1]. Due to the intrinsic relationships existing in multivariate data, traditional data analysis methods such as classification and clustering are having difficulties while discovering efficient relationships between several variable simultaneously. As the database size increases dynamically and the dramatic changes in the use of data will show adequate results on predicting the effect on change in one variable will have on other variable. This introduces the necessity of Multivariate data analysis.

Multivariate data clustering takes A larger group of objects and measurements on them of some properties. on the basis of these, multivariate data clustering then attempts to group the samples on the basis of samples similitude or dissimilarity. The analysis of clusters aims to systematize variable information to create relatively homogeneous groups or clusters. Very homogeneous internal and highly external homogenous This method is used to form clusters. A number of similarity measures can be defined in order to recover data elements for specific classes, corresponding to data nature, rationale and means governing cluster formation. The analysis of the cluster is an iterative optimization process and differs from the automatic task and involves knowledge discovery and interaction with multi-target testing [8]. Implementation is the prerequisite for inspecting the similarity. It is an important task in the analysis of statistical data, exploratory data mining, used in many fields, together with machine learning with the collection of information, pattern recognition and image analysis and bioinformatics. With the developments of multivariate data, more attention has been paid in identifying predictors or independent variables and dependent variables. Most of the research work was concentrated on identifying only dependent variables rather than predictor variables containing one or two datasets.

However, those existing work mostly considered identifying dependent variables using univariate data analysis. Furthermore, there is a large difference between data predictors in different datasets, primarily because of the diverse changes in multivariate data and variables. While univariable approaches are intended to research and understand systems, when more complex systems are analyzed, they tend to fail. Thus for multivariate data analysis, the following challenges remain: How should the missing data be addressed? How should the over simplistic and overoptimistic assessment of the data is handled? How can the relations between the variables studied be detected? How to know the *covariance* or *correlation* that exists with Multivariate data [12]. To solve these problems, the proposed work will concentrate on identifying dependent and predictor variables in presence of missing data and discovering correlation and eccentric relationships among these variables.

In this paper, we study an improved multivariate data

Revised Manuscript Received on February 14, 2019.

M. Sankara Prasanna Kumar, Research Scholar, JNTUH-Hyderabad, Assistant Professor, AITS Rajampet, A.P., India. (sankaraprasannakumar@gmail.com)

A. P. Siva Kumar, Assistant Professor, Department of Computer Science and Engineering, JNTUCE, Anantapuramu, A.P., India. (sivakumar.ap@gmail.com)

K. Prasanna, Associate Professor, Department of Information Technology, AITS, Rajampet, A.P., India. (prasanna.k642@gmail.com)

clustering using heuristic data intensive computing and relevance feedback learning approach MRF that is specially designed to discover very large Clusters over multivariate datasets is described. MRF implementation makes use of a heuristic data intensive computing (HDIC) and efficiently reduces the data preprocessing time over the dataset. Clustering the data is performed with proposed Relevance feedback learning algorithm with consensus distance functions. These functions are selected as the change in the cluster ensemble selection, combine and reduction. In this proposed approach we have used a new kind of distance functions such as Camberra, Chi-square and Cityblock. The empirical analysis shows that better cluster sets of various multivariate datasets are achieved with the proposed approach. The experimental findings show that this approach results better in the mining and operations with KNN in different settings of multivariate datasets.

The paper is organized as follows. Section 2 introduces basic preliminaries and definitions associated with multivariate data. Section 3 defines an approximation of heuristic data intensive computing for multivariate data variable. The relevance based learning algorithm using consensus functions used to evaluate cluster ensemble selection, combine and reduction are presented in section 3. Comparisons with existing methods on simulated and real datasets are presented in Section 4, and a discussion concludes the paper in Section 5.

II. BASIC PRELIMINARIES

Multivariate analysis consists of analysis of data sets with more than one variable. This section provides a non-exhaustive description of multivariate data and set of definitions and techniques associated with a given multivariate data. In reality, univariate analysis often leads to incorrect interpretations due to more than one variable in the dataset. Compared to univariate analysis, Multivariate analysis provides highly correlated behavior of data and depicts more accurate conclusions.

A. Definitions

Basically there are four types of multivariate data. Many data sets can fit into one of them.

Definition 1: *A single sample with several variables measured on each sampling unit.*

Definition 2: *A single sample with two sets of variables measured on each unit.*

Definition 3: *Two samples with several variables measured on each unit.*

Definition 4: *Three or more samples with several variables measured on each unit.*

There are plentiful analysis tools used for multivariate data in real life applications ranging from exploratory data analysis to quantitative regression models [6]. To classify samples into groups with similar characteristics, there are a variety of classification models such as Soft Independent Modeling by Class Analogy (SIMCA) or Discriminate Analysis [6]. One thing to keep in mind is that modeling is an iterative process. Depending on the objective, multivariate data can be used to comprehend and model several outcomes. This section will explore some of the basic techniques used in multivariate data analysis.

Principal Component Analysis is the most versatile method used for multivariate data [8]. The objective of PCA is to break the data table into a new set of unrelated variables with the appropriate measures [8]. These variables are called, depending upon the context, principal components, factors, eigenvectors, singular vectors, or loadings. Correspondence Analysis (CA) is a simplification of PCA to contingency tables. The orthogonal decomposition of factors associated with table is done using the Chi-square values. The CA has a symmetric role in rows and columns of the table. CA is generalized as a Multiple Correspondence Analysis if several nominal variables are analyzed. [15]. Multidimensional scaling (MDS) clustering technique is applied when the rows and the columns of the data table symbolize the identical units and when the measure is a distance or a similarity [2].

K-Nearest Neighbour derives expertise from previous data analysis training patterns. By using majority votes among the nearest K neighbours, the input test data are divided into a certain class. The nearest K neighbors will be selected using a pre-defined distance measurement (hamming, euclidean distance etc.). The next training pattern is sequentially calculated and selected as a neighbor to sample selected measures [17]. This is used consecutively for calculating and selecting the closest training pattern to provide a sample of the actions selected.

III. MRF: MULTIVARIATE DATA ANALYSIS USING RELEVANCE FEEDBACK LEARNING

A. Heuristic Data Intensive Computing

In the knowledge extraction process, the MRF calculates the distances between the samples and near data values. The input data set properties will influence the similarity measures [4]. If the data contains high variability due to noise, in the reduced data variation will not be represented properly. A data intensive computing step is desirable proceeding to MRF approach to get rid of the i due to issues related to noise, outliers and variable boundaries.

1) Boundary value analysis and Outliers

Boundary values or missing value and outliers are the variables which will affect the mining process. The best way to treat boundary values is to delete or skip the corresponding row and column of the data table which contains boundary values or missing values. Outliers are the measurement values recorded in variables outside the data limit values. [10].

2) Consensus Distance functions

The MRF approach starts with calculating the distances or similarities among the set covariance samples in the multivariate dataset. For Multivariate data analysis, the commonly used distance measure is Euclidean distance. It is sensitive to outliers; because small differences are weighted heavily compared to large differences. In multivariate data,

Euclidean distance observes zero value when two samples

observed concurrently. Due to these reasons, in this paper we have used other consensus distance functions including variants of ED such as, Camberra distance, Chi-square distance, City-block (Manhattan) distance [10].

The feedback learning for relevance is selected using the preset consensus distance measures in order to detect the chosen metric. The similarities with relevant feedback are as follows:

a) Camberra Distance

The Camberra distance *d* between objects x and y in an *m*-dimensional object space is given by:

$$D(x,y)=\sum_{i=1}^{m}\left|\frac{(x_i-y_i)}{(x_i+y_i)}\right|$$

Where $y = (y_1,y_2,y_3,..., y_m)$ and $x = (x_1,x_2,x_3,..., x_m)$ are two points in Camberra dimensional space of m.

b) Chi-Square Distance

The chi-square distance *d* between objects x and y in an *m*-dimensional object space is represented as

$$D(x,y)=\sum_{i=1}^{m}\left(\frac{1}{sum_i}\right)*\left[\left(\frac{x_i}{size_x}\right)-\left(\frac{y_i}{size_y}\right)\right]^2$$

Here sum of tuple values as sum_i for attribute i taking place in the training dataset, and $size_x$ is the sum of all values in the object x. In addition, qualitative variables can be altered into binary variables in multivariate analyses [5].

C) Cityblock Distance

The well-known distance city block also called a Manhattan similarity measure may be perceived as

$$D(x,y)=\sum_{i=1}^{m}|(x_i-y_i)|$$

Where $x = (x_1,x_2,x_3,..., x_m)$ and $y = (y_1,y_2,y_3,..., y_m)$ are two points in Cityblock dimensional space of m.

d) Accuracy

Accuracy should be measured by taking true positive and true negative aspects into account.

e) F1-measure

Unlike accuracy, takes into account the prediction of classes separately.

The f1-measure with respect to the positive class is based on the confusion Matrix and is defined as:

$$F1\ Measure(F1-M)=\frac{2TP}{(2TP+FP+FN)}$$

Where TP: True Positives, FP: False Positives, FN: False Negatives

B. Relevance Feedback Learning

Relevance feedback learning is a characteristic of some information collection systems. Feedback connotations are used to get user vector values and to analyze the data to ensure that the results are relevant to a new query [13],[16]. Feedback on relevance supports the quadratic distance metrics in general. It makes interactive learning easier to refine results. The similarity of objects may also be described as the degree of relevance between pair - wise objects and the highest grade combining with data set dimensions. [11].

One of the solutions taken was to directly derive the benefits of human expertise from the so-called relevance feedback (RF) recovery results. The user has to provide feedback for some retrieval objects by finding the results as data objects, whether they are relevant or not. The system then calculates repeatedly, using this information, an improved representation of the information requested and also refines the recovery.

One of the first to thrive feedback techniques for relevance was Rocchio's technology. The classical approach of the ranking reflects the hypothesis that tuples are distributed uniformly, but through the distribution of the class label each tuple has the same relevance. A score of the data set stimulates the comprehensive performance of the technique. In the assessment of the tuple distribution in datasets, the ranking score can also be considered. In view of their optimistic and pessionist tuples, the object is updated to adjust the location of the original object in n-dimensional space with its associated vital factors. Feature relevance assessment (FRE) is another example when the user can consider certain precise objects more important than others in a given query. Every object has a meaningful weight that has improved variance in value compared to objects with smaller variations. Machine learning techniques have recently initiated an important feedback approach.

Relevance Feedbak (RFB) is a technique used in non - interactive experimental environments which helps searches to increase the accuracy of the query statements.

1) Pseudo-Code of RFL Technique

RFB pseudo-code can be represented as:

a) Initially, the user in the clustering stage sets the constant 'k' and learning parameter (ℓ).
b) Exploratory search is conducted by clustering a data point unmarked by calculating the distance between the base point and all trail points in the dataset.
c) During exploratory search the new centroid is computed $X_c = \frac{1}{N}$
d) The distances obtained will then be sorted by base point 'k', and the nearest neighbors will be shown in the graph.
e) Repeat exploratory search (step 2) using Xc as new base point.
f) The distances between the base point and user vector points of the cluster shall be calculated.
g) The neighbor is considered to be genuinely positive if he

is in a given class, otherwise it is considered to be true negativity. Precision and F1 measurement is based on the true positive and negative value.

Relevance feedback (RFB) is a technique that helps searchers to improve the accuracy of their query statements and has been shown to be effective in non-interactive experimental environments [9].

IV. PERFORMANCE ANALYSIS & RESULTS

A. Performance Evaluation of different Datasets using RFB Technique

Experimental results are shown in this section are to study the impact of similarity metrics such as Camberra distance, Chi-square distance and Cityblock distance are analyzed for the number of nearest neighbors from different classes of dataset for a particular seed point [7]. 5 datasets having different characteristics have been chosen from the UCI repository [14] to observe the impact of using diverse similarity measures on them. To be able to convert the relational datasets into transactional datasets, all numeric attributes are discretized. The same entropy-based discretization method [3] used in CBA [7] is used to categorize the continues attributes. This method is a supervised top-down approach which discretize the values with no parameters. The datasets and some of their properties and characteristics are shown in Table 4.1.

Table 4.1: Dataset characteristics

Dataset	Data type	Attribute type	No. of Instance	No. of Attributes	Class
IRIS	Multivariate	Real	150	5	3
Wine	Multivariate	Real	178	13	5
Zoo	Multivariate	Categorical	101	17	7
Glass	Multivariate	Categorical	214	10	6
Breast Cancer	Multivariate	Integer	699	11	2

Table 4.2: Accuracy of all datasets for different distance measures based on RFB technique with k = 3.

Dataset	Camberra	Chi-square	Cityblock
IRIS	66.666	33.333	66.666
Wine	38.000	34.333	35.000
Zoo	43.333	43.333	43.666
Glass	13.333	13.333	43.666
Breast Cancer	66.666	33.333	66.666

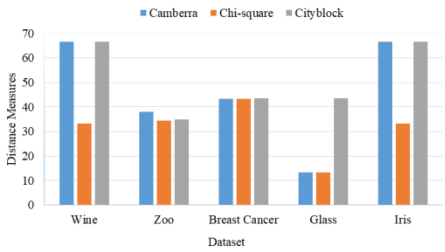

Figure 4.1: Accuracy Graph

In Table 4.2 Camberra, cityblock distances, is better accurate on iris data set than other similarities. For camberra and city block distances the accuracy is better in wine data set. The camberra distance is improved in the zoo dataset. City blocks and lens data are better for lung cancer, for the distances from city blocks. If the data set is very large and in addition the wrong assumption of the seed point results in a higher overhead for the user to find with possible lengthy reaction times. Precision in comparisons to the KNN algorithm is increased, which demonstrates the increase in number of true positive.

Table 4.3: Comparing the change in Chi-Square similarity during selection, combine and reduction phase of a cluster with k = 3.

Dataset	% change in selection	% change in combine	% change in reduction
IRIS	1.57	0.81	1.57
Wine	22.56	17.36	22.56
Zoo	19.44	19.49	19.49
Glass	14.78	8.41	7.66
Breast Cancer	2.05	0.48	-4.96

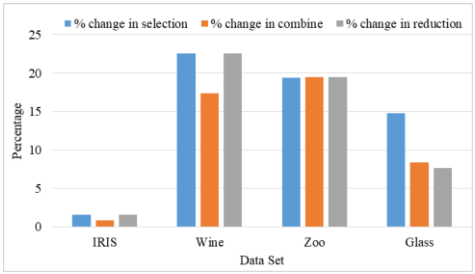

Figure 4.2: Comparison Graph

Table 4.4: Change in Cluster reduction while using Relevance Feedback Learning with k=3.

Dataset	Cluster reduction	F_1 Measure		Accuracy Measure	
		KNN	RBL	KNN	RBL
IRIS	91.58	91.73	94.48	92.00	94.67
Wine	99.60	74.79	92.51	77.58	92.60
Zoo	99.77	66.00	91.26	81.16	94.99
Glass	86.65	53.08	50.64	69.82	63.19
Breast Cancer	96.01	94.22	95.74	94.84	96.13

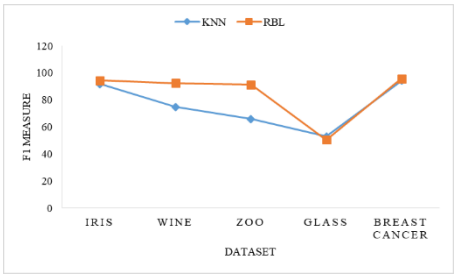

Figure 4.3: Dataset Vs F1 Measure Graph

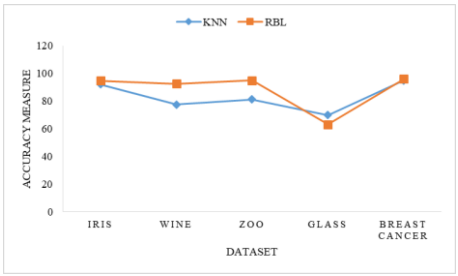

Figure 4.4: Dataset Vs Accuracy Measure Graph

Table 4.5: Performance comparison of MRF Learning with KNN with respect to time in milliseconds k=3.

Dataset	KNN	RF Learning
IRIS	0.069	0.049
Wine	0.099	0.059
Zoo	0.097	0.092
Glass	0.082	0.045
Breast Cancer	0.095	0.049

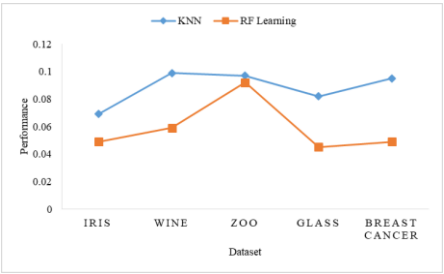

Figure 4.5: Dataset Vs Accuracy Measure Graph

V. CONCLUSION

In this paper, an improved relevance feedback learning algorithm MRF is presented to discover clusters over multivariate datasets. In our algorithm we iteratively constructed cluster ensembles using consensus functions. These functions are selected as the change in the cluster ensemble selection, combine and reduction. In this proposed approach we have used a new kind of distance functions such as Camberra, Chi-square, Cityblock. Similiarity measures are analyzed for prominent KNN and MRF UCI real - life data sets and comparable results.

As the number of clusters increase, data objects decrease in more classes with the KNN technique. This technology can not therefore be used for few data sets, since there is no change in accuracy analysis when the number of neighbours increases. Therefore, the selection of threshold parameters with KNN technique is difficult before the technique is run and is always based on a value of k for clustering.

The empirical study shows that in different environments our approach has achieved great mining efficiency. Furthermore, we demonstrate that in the discovery of clusters this algorithm also achieves the highest accuracy and the best F1 measure compared to previously developed algorithms..

REFERENCES

1. A. Rencher "Multivariate Statistical Inference and Applications" Wiley series in probability and statistics: Texts and references section. Wiley, 1998.
2. Borg I., & Groenen P. "Modern multidimensional scaling" New York: Springer-Verlag, 1997.
3. Chen, Ching-Yi. and Ye, Fun., "Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis" IEEE ICNSC 2004, Taipei, Taiwan, R.O.C., Vol-,No-,2004,pp. 789-794.
4. Das, S., Abraham, A. & Konar, A. "Metaheuristic pattern clustering – an overview" Metaheuristic Clustering SE – 1, Berlin, Germany: Springer-Verlag, 178, 1–62, 2009.
5. Ellison GN, Gotelli N "A primer of ecological statistics" Sinauer, Sunderland, MA, 2004.
6. G. Gan, C. Ma, and J. Wu "Data Clustering: Theory, Algorithms and Applications" volume 20 of ASA-SIAM Series on Statistics and Applied Probability, pages 1–466. 2007.
7. G. Giacinto "A Nearest-Neighbor Approach to Relevance Feedback in Content-Based Image Retrieval", ACM Int. Conf. on Image and Video Retrieval, Vol-,No-, 2007.
8. Jain, A. K., Murty, M. N., & Flynn, P. J. "Data clustering: A review" ACM Computing Surveys, Vol-31, 1999, pp.264–323.
9. Niall Rooney "A relevance feedback mechanism for cluster-based retrieval, Information Processing and Management" an International Journal archive Volume 42 Issue 5, 2006, pp.1176 – 1184.
10. Osborne JW, Overbay A "The power of outliers (and why researchers should always check for them)" Pract Assess Res Eval 9:1–12, 2004.
11. S.H. Huang, Q.J Wu, and S.H. Lu "Improved AdaBoost-Based Image Retrieval with Relevance Feedback via Paired Feature Learning" ACM Multimedia Systems,Vol-12,No-1,2006,pp. 14-26.
12. T. J. Fisher and X. Sun "Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix" COMPUTATIONAL STATISTICS & DATA ANALYSIS, 55(5):1909–1918, MAY 1 2011.
13. Vinay, V., Cox, I. J., Milic-Frayling, N., & Wood. K "Evaluating relevance feedback algorithms for searching on small displays" In D. E. Losada & J. M. Ferna´ndez-Luna (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science (LNCS) Berlin: Springer.2005,pp. 185–199.
14. UCI Repository for Machine Learning Databases retrieved from the World Wide Web: http://www.ics.uci.edu/~mlearn/MLRepository.htm,
15. Weller S.C., & Romney A.K. "Metric scaling: Correspondence analysis" Thousand Oaks (CA): Sage, 1990.
16. Xiang Sean Zhou, Thomas S. Huang "Relevance feedback in image retrieval: A comprehensive review" Proc. Springer-Verlag on Multimedia Systems, vol 8,2003,pp. 536–544.
17. Xin Huang,Shijia Zhang,Guoping Wang and Heng Wang "Optimal Matching of Images Using Combined Color Feature and Spatial Feature" Springer-Verlag Berlin Heidelberg,Vol-3991,2006,pp-411-418.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication