Clustering of the Multi-Value Documents based on Probabilistic Features Association Mechanism

P Gopala Krishna, D Lalitha Bhaskari

Abstract: ***It is becoming increasingly difficult to cluster multi-valued data in data mining because of the multiple data interval values of individual functions. Identifying a clustering model that is appropriate for these disguised multi-valued data deployments in data analysis applications is an open problem. To answer this question, this paper proposes a feature selection based on the probabilistic features association mechanism (PFAM). The problem is mainly due to the difficulty in identifying the class information and the multiple values for each individual features. This work explores the problem of unsupervised feature selection through computing the probabilistic association score and multi-value data reformation for effective clustering in multivariate datasets. By minimizing a reformation clustering error, it can conserve together the degree of similarity and the categorization information of the actual data contents. The proposed approach is evaluated the clustering purity and Normalized Mutual Information on multivariate document datasets. The experimental evaluation shows the improvisation of the proposed approach.***

Keywords*: *Feature selection, Probability Association, Clustering, Multi-value document

I. INTRODUCTION

Clustering is a computationally efficient and accurate method for data mining for data classification. Large data sets must be processed in terms of the amount, size, and complexity of data classification. Many different approaches have been proposed in [3], [5] to support efficient and accurate clustering. In general, uncertain object data distributions can be expressed by distribution probabilities [15],[19],[37]. The difficulty of clustering multi features objects according to the probability distribution occurring in many situations. The most effective medium is to describe this feature, attribute selection [11], [12], [22] and feature extraction [21], [27],[28] which can minimize the dimensionality of meaningful feature finding in a set of features of the objects.

The feature selection methods are often categorized as "supervised" or "un-supervised" methods for the regulation of feature selection. The supervised feature selection method [24], [29] uses the relationship between the characteristics and the label information to lead the assortment of features that are significant and relevant. So researching the great value of its analysis will make it difficult to choose features that are not being monitored and use as much data as possible. This paper focuses specifically on non-supervised feature selection issues that occur because the information that references feature selections is not labeled.

The current unsupervised function selection algorithm [5], [8] has been widely used for the clustering of textual information. In the collection of text, text documents or words are always expressed in word bags that cause a high-dimensional space. The method of selecting the unsupervised function selects a subset of words from the word of the actual data space usually according to certain condition [20], [22], [25], [30]. There are two types of unsupervised selection algorithms [1], [8], [9], [18] that maintain the similarity function and the maximum clustering efficiency. The methods selected to maintain the characteristics representative of similarity that finest preserve the traditional structure of the unique data space. For instance, if the points are near the data point of the data allocation node, these data points must be considered near to each other based on the selected characteristics. On the other hand, maximizing the embodiment of operational approaches can distinguish and decided features that can be used to develop a grouping criterion. As Tang et al. [32] and Yang et al. [6] uses the conceptual label menu to select the functions that can be broken to increase the effectiveness of data point clustering.

In this article, we investigate the difficulty of selecting unsupervised functions from the point of view of the probabilistic feature association mechanism (PFAM) and multi-valued data reform. The PFAM proposal for selecting data point features will adjust the natural quality measurement for the selected feature to be approximated by linear combinations and selected characteristics data. The scarcity of the function selection matrix can reduce function noise and duplicate data from multi-valued data. We will evaluate an extensive range of experimental data clusters in the multi-value data set, to evaluate the effect of the proposed method.

The rest of the paper organized in the following mode. Section-2 presents the literature review of the feature significance and clustering,

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author

P Gopala Krishna*, Associate Professor, Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India E-mail: gopalakrishna.aucsse@gmail.com

D Lalitha Bhaskari, Professor, Department of CS & SE, Andhra University College of Engineering (A), Andhra University, Visakhapatnam, India. E-mail: lalithabhaskari@yahoo.co.in

Section-3 present the proposed probabilistic features association mechanism, Section-4 presents experiment dataset and measures, Section-5 discuss the result outcomes and Section-6 conclude the conclusion of the paper.

II. LITERATURE REVIEW

Often the number of variables or characteristics increases to account for information in many domains due to other types of data, such as images, text documents, medical information, and other information extraction [1]. Sometimes this high-level data consists of multiple values based on observations of the characteristics. In fact, all functions are often unrelated and uniquely characterized, often because they are related or overlapping, and are sometimes very loud for choice [5], [8]. Their capabilities are in higher dimensional feature through existing learning models that are hindered by excessive adaptability, small effectiveness, and worst performance [1]. It is as a result difficult to learn how to improve the accuracy and understanding of results, so it should eliminate unnecessary and repetitive attributes that contain large amounts of data that it must select from a subset of features.

San et al. [13], try to reduce dependence on the introduction of a new "cluster center" concept for the categorical objects called representatives. It specifically defines the use of a representative cluster to map the distribution of the categories values shown in the cluster. The measurement of the necessity between the representatives of the materials and cluster that are unconditionally defined in support of the relative frequencies of values in the category and that there is only a coincidence between the category and the cluster values. As a result of the cluster formation approach as a *k*-representative algorithm. In [6], the *k*-representative algorithm showed an effective cluster mechanism for the data categorization.

In particular, the *k*-modes method [2], [4], and [7] algorithms are used primarily to make a simple proportional-based content comparison and instead of the method and it has discovered the potential of the cluster. This feature defines a feature as a data point in a cluster, which is the most common value of all values found in clusters assigned to a set of domains. It can also be combined with the *k*-mean algorithm to process and collect large-scale data series of different scopes using mixed numerical databases and blending categories. It should be noted that although a cluster can have more than one method in a *k*-mode cluster, the algorithm relies heavily on method selection for cluster processes [38].

A. The significance of Feature Selection

The process of identifying appropriate effective features is an important task and one of the most widely used mechanisms in various domain analysis and information mining, and the methods for selecting various features for machine learning applications have been studied and proposed [2], [7] [9], [10]. According to the methods using labels information can be divided into supervised algorithms [5], [18], [20], semi-supervised algorithms [27], [33] and unsupervised algorithms [18], [33]. The supervised methods are examined in such a way that they can select distinctive features because the data is encoded in the distinctive label identifiers. With the available feature correlation, one can expect sparse-based methods that have been studied in [18], [26] for the feature relation learning. However, least label marked data and large label unmarked data can easily be configured to build a common data set. But the problem of describing with the least label marked problem generates a challenging problem for the supervising algorithm[26].

Since there is not enough information about data labels and small labels, supervisory algorithms often fail to remove many specific features by mistake or by using one of the irrelevant selection features. Therefore, the supervisory function selection is developed simultaneously to utilize unlabeled label data. Without a label to guide the search for a unique feature, the selection of an unchecked feature appears to be a much more difficult problem [36] that assesses the nature of the interest in the ability to maintain certain attribute of an element. In several real-world applications, the lack of unlabeled data and the rapid accumulation of higher dimensions confront the cost of label identification. As a result, a very promising and autonomous development of unsupervised feature selection techniques [5], [6], [20], [25] are required.

B. Feature Selection in Clustering

Unsupervised clustering based on a feature selection suffers as a result of the lacks of label information that identifies a subset of characteristics of having a unique cluster in accordance with the specified standard for the more difficult clustering criterion [32], [36]. Clustering-based feature selection methods use functional concepts to create virtual labels, perform feature selection, and produce pseudo-labels for data occurrence.

Z. Li et al. [16] describes the function of feature selection performed on pseudo-label information used for spectral clustering of data being performed simultaneously for all instances. Yang et al. [6] suggest that an integrated framework for joint class structure for linear classifier determination and autonomous feature selection for input instances of data for differential analysis under the assumption of labels can be predicted. E. Bradley et al. [36] suggest the cluster-based quality and immediate Tang et al feature-based assumptions. [32] also performs pseudo-label generation to incorporate discriminate analysis for unsupervised feature selection.

Clustering learning algorithm [15], [27], [28] are generally used for unsupervised feature selection. The wrapper method [18], [32], [36] uses a predictive model for a subset of the scoring function in case of unsupervised feature selection. It will be used to train the small test set with each new model as subsets. It takes the advantage of learning outcomes for feature selection using feature learning algorithms.

A.E. Brodley et al. [36] explored the maximum likelihood separation and distribution for feature selection, clustering, Gaussian clustering, and sequential search. An objective method for optimization for "least squares" is described in the "Q-α algorithm" [26]. It measures the spectral characteristics of the input data points of the cluster capacity and analyzes them through affinity matrix analysis. The MCFS [20] uses spectrum analysis to measure the correlation between various functions. The feature selection procedure is depended on the least normalized square optimization problem.

The complexity of unsupervised feature selection is mainly solved by using probability mechanisms [32], [34]. These methods are classified into category tags based on feature selection based clustering that considers latent variables. S. Boutemedjet et al. [23] proposed a visual feature and used a separate class feature to generate a clustering model. Y. Guan et al. [31] proposed a probabilistic model of global integration capability and unsupervised feature selection. W. Fan et al. [17] proposes a framework for inferring changes in the "unsupervised non-Gaussian" approach to feature selection. In the earlier researches, the proposed work was based on the unsupervised probabilistic feature association mechanism of data reform and feature selection and formulated this problem. This paper will maintain similarities between the two and will choose the best features to distinguish the unclassified data information using a probabilistic features association for enhancement in the clustering mechanism.

III. PROPOSED FEATURES ASSOCIATION MECHANISM

A. Problem

Cluster analysis is a method of exploring data for the purpose to collect a set of objects inhomogeneous clustering so that the elements in the cluster should be highly similar [21]. However, recent advances in clustering technology represented by a general vector of quantitative values in a multidimensional space in the database. It is now usually recorded as a data or weight probability distribution [18]. Classical multivariate data analysis and similarity distance between data play an important role. The other distances measure technology is according to the type of measurement to be selected. Although distances to various similarities are defined in existing data analysis supporting the environment of the object and data, the following suggestions have been made to analyze the histogram or interval data. However, there are many scenarios where the interval data value is most appropriate for real data.

In general, data interval values cannot be described by single-valued variables. Some data sets first to contain the interval attribute. For example, the attribute age will be recorded as an interval, such as [0, 10], [30, 40], and so on. One aspect of the number of instances of a data set is affected by many attributes, which are primarily subject to scalability issues. Researchers and practitioners in all fields are experimenting with automated methods to analyze data, preferably while maintaining important information while reducing the size of the data. Therefore, in order to perform unsupervised feature selection, we will determine that for a given data matrix *X* and matrix feature *F*, we will determine the feature collection of *S* of size *k* from among the *n* features to define the problem in order to identify the information in the actual data.

B. Probabilistic Features Association Mechanism

The well-known feature extraction algorithms, known as "principal component analysis (PCA)", extract differential characteristics depend on information acquisition. Motivated by the initiative of the "PCA algorithm", we take advantage of the probability-feature association mechanism (PFAM) based on the association of data and suggest new feature selection criteria. It first introduces the data feature regeneration condition from the perspective of data characteristics, and then provides PFAM reconstruction particulars of the data for the data points in the source data space.

The mechanism of feature relevance describes the relationship of individual features that identify the beneficial impact of each feature on each other. For any given set of data, we applied the "*k*-means clustering algorithm" to acquire the primary number of clusters. The best results for the "*k*-means algorithm" and objective function applied to different starting points are recorded for feature selection as shown in Fig. 1.

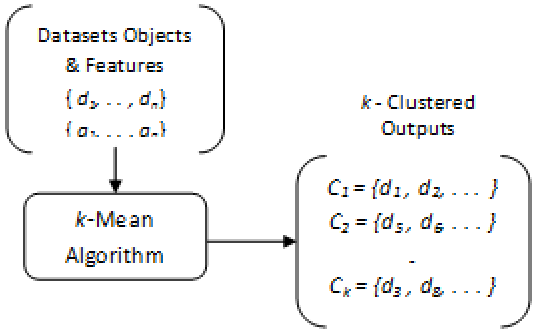

Fig.1: Construction of initial *k*-Cluster results

Let us suppose it as a graph G form, where there are a series of edges associated features that are shown as G = (*N*, *E*), where *N* is the set of nodes and the number of nodes *n* = | N | and *E* is the set of edges of the vertex of the graph. We consider the main purpose of the dataset as *N* and the number of features such as *E*.

We consider that a data source is composed of *F* features of the size *a*, and objects of a unique class of a data class as *U* of counting *n*, which can be shown as $n \epsilon N(G)$ and $a \epsilon E(G)$. There are two vertices $\{ n_i , n_j \}$ adjacent, if there are any adjacent edges features share a common end-vertex then it shown as $\{ a_1 , a_2 , . . . , a_d \}$.

Suppose, $N (G) = \{n_1, n_2,...,n_n \}$, then an, n × n, *(0,1)* matrix as, $E := E(G)=(a_{ij})$, is called the association matrix of *G*, in such case if, $a_{ij} =1$, then $n_1\ n_2$ $E\ (G)$, else $a_{ij} =0$. In order to construct an association matrix *E* on the graph, we consider the feature weights as *0* and, where the weight $a_{ij} =1$, if and as long as nodes *i* and *j* are associated by a feature edge as shown in Fig.2.

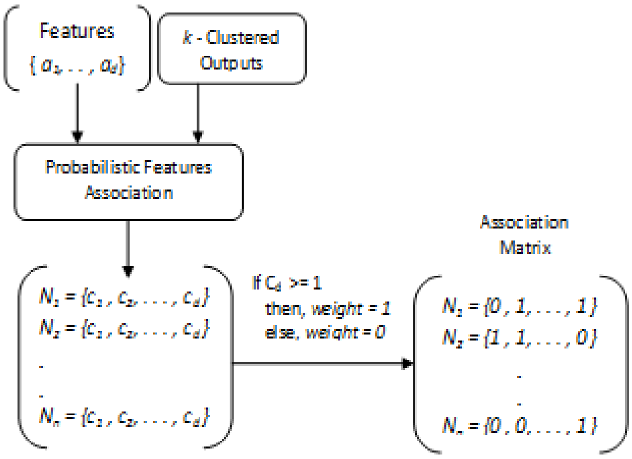
Features
{ a_1, . . , a_d}
k - Clustered
Outputs
Probabilistic Features
Association
Association
Matrix
N_1 = {c_1 , c_2, . . . , c_d}
N_2 = {c_1 , c_2, . . . , c_d}
N_n = {c_1 , c_2, . . . , c_d}
If C_d >= 1
then, weight = 1
else, weight = 0
N_1 = {0 , 1, . . . , 1 }
N_2 = {1 , 1, . . . , 0 }
N_n = {0 , 0, . . . , 1 }

Fig.2: PFAM Association Matrix Construction

DOI: 10.35940/ijitee.A4538.119119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Note that *E (G) = A*, where *A* is indexed by 1 in the matrix if the feature weight element count is $c_d > 1$ and 0 if it has the feature weight element count is $c_d = 1$.

The likelihood probability association's (PA) and the association matrices are measured by *PA = (D - E)*, where *"D* is a diagonal matrix" represent as "$D(i,\ i) = \sum_{1}^{d} A(i,d)$".

Since we consider two different values of the same object class with the characteristics of building the matrix of the association, *D* is likely to have a total value of 0 because they are not connected. So, in such case *PA = E (G).*

Based on every single feature implication of the *PA = E (G)* function, the necessary essential feature characteristics for clustering were identified as illustrated in Fig. 3.

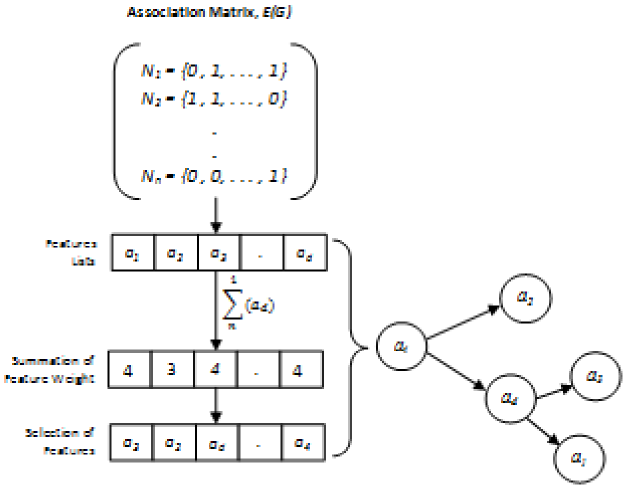

Fig.3: Feature Selection and relation association graph generation

Let it assumed, if a feature collection "*F = {a1, a2, a3, a4, a5}"* and the product found on the weight of a feature such as, *W = {4, 3, 5, 1, 4}*. So, based on the list is the essential element *W* for cluster, *Z = {a4, a2, a1, a5, a3}* and association relation graph is generated. Based on graph related to the features, we will assess the effectiveness of a cluster for choosing a different number feature. The following sections were selected, through the number of features set to solve the multi-value datasets and evaluate the method of cluster approach to measure our approach as discussed in [4].

IV. EXPERIMENT DATASET AND MEASURE

In this section, we present the evaluation impact of the proposed PFAM approach to a categorical dataset on the chosen features of the independent and non-asset dataset. The evaluation data set is collected from the real-time UCI data store. The elements of each dataset are concise in Table-1. For this dataset, each of the properties of their public availability category selected for testing in our algorithm can be tested.

TABLE I
UCI CATEGORIZED DATASETS

Datasets Type	No. of Class	No. of Features	No. of Instances
Car	4	6	1728
Mushroom	2	22	8124
Nursery	5	8	12960

Evaluating the quality of clustering is often subjective and difficult [27]. In order to achieve high similarities within clusters and clusters, clustering objective functions with low intimacy are generally intentionally designed. This can be seen as an interior standard for the quality of the cluster. Nevertheless, as it is in the literature, it is not necessary to translate the good effects of an application into a good score for internal references.

Here, we use two criteria: "*Purity*" and "*Normalized Mutual Information (NMI)"* as the assessment condition for the results in the same way as [14]. This method is used for clusters assigned to assess how well the objects in the cluster are matched to the basic information of the actual class.

The clusters represent with $C = \{C_1 ,...,C_J\}$ to the dataset generated by the cluster algorithm and division given by the unique classification partitioned, $P = \{P_1,..P_I\}$. The *"J"* and *"I"* are the number of clusters marked with |C| and the number of cluster classes mentioned by |*P*|, whereas *N* shows the overall number of data subjects in the dataset.

- *Purity Measure:* An evaluation of purity measure is simple and transparent. Each individual cluster is assigned to a class divided by the precision of the specified object, divided by the number containing the number of objects within this exact allocation, and most often the purity is computed in the cluster dataset [4]. If the purity is high the improved the clustering. It is computed using the equation-1 as given below.

$$Purity(C,P) = \frac{1}{N}\sum_{j}\max_{i}|C_j \cap P_i| \qquad (1)$$

- *NMI Measure:* The NMI metric provides several independent pieces of information in the cluster [8]. When this clustering completely matches the actual partition, the measured value has the maximum value [4]. The NMI's average common information is calculated among a pair of clusters and individual classes using the equation-2.

$$NMI(C,P) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}|C_j \cap P_i|\log\frac{N|C_j \cap P_i|}{|C_j||P_i|}}{\sqrt{\sum_{j=1}^{J}|C_j|\log\frac{|C_j|}{N}\sum_{i=1}^{I}|P_i|\log\frac{|P_i|}{N}}} \qquad (2)$$

Many previous studies have been used only to analyze purity metrics to evaluate cluster algorithm performance. However, if more clusters are available, purity measurements can be easily performed. Especially if the data purity Purities has to measure everything as 1. Also, many departments have the same purity. For example, the object data for each cluster is different from the other clusters by number. The number of items that the cluster performs Thus we measure the excellent efficiency of the NMI and how the cluster results are equivalent to the actual class.

V. RESULT ANALYSIS

In this section of the section, we present an experiment to compare the cluster performance between "*k*-mode", "*k*-representatives" [8], "*k*-representatives-Modified" [4] and proposed PFAM.

We start that the constraint *k* is equipped to the number of data classes and execute our approach across the two different databases designed to measure performance and metrics. The average run we have conducted to assess the performance of the two metrics in different datasets.

The results in Tables 2 and 3 show a series of category categories considered purity and NMI results.

TABLE III
COMPARISON OF PURITY RESULTS

Datasets Type	PFAM	k-mode	*k*-representatives	*k*-representatives -Modified
Car	0.822	0.816	0.715	0.721
Mushroom	0.918	0.891	0.859	0.894
Nursery	0.615	0.481	0.481	0.485

TABLE IIIII
COMPARISON OF NMI RESULTS

Datasets Type	PFAM	k-mode	*k*-representatives	*k*-representatives -Modfied
Car	0.228	0.081	0.075	0.142
Mushroom	0.641	0.541	0.479	0.541
Nursery	0.127	0.051	0.481	0.073

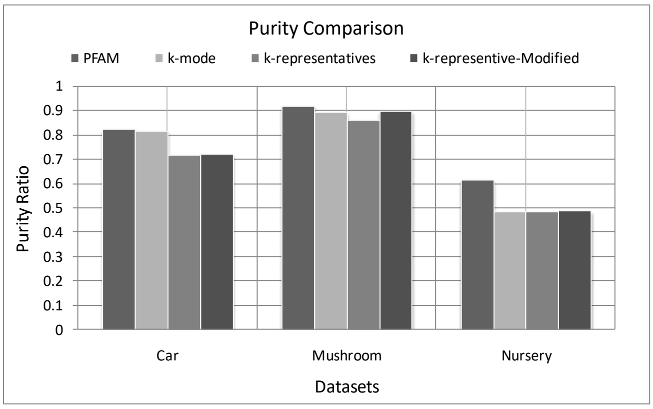

Fig. 4 Datasets Purity Result Comparison

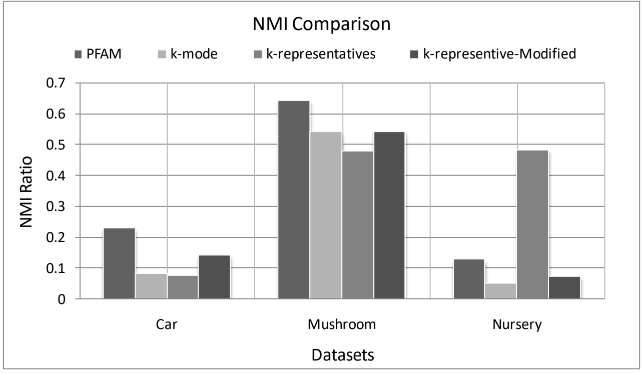

Fig. 5 Datasets NMI Result Comparison

PFAM has very good results for each data set compared to *k*-mode and *k*-representative. When comparing performance against *k*-representative-modified in the proposed PFAM approach, especially in NMI values, the results are enhanced in each case. Finally, the new approach developed represents the performance in comparison to algorithms similar to the previously developed to cluster data. Fig. 4 and 5 show the comparison result by comparing the purity and NMI ratio.

VI. CONCLUSION

In this paper, we investigate an unsupervised feature selection problem for multi-valued datasets that reconstruct data to use cluster through the use of a similar model of probability. Our approach solves the difficulties of choosing a feature in unsupervised data through the easy integration of data restructuring and the Probabilistic Feature Association mechanism into a general mechanism. The proposed approach shows the ability to determine the information in the actual data space by keeping the highest similarities through minimizing graph error and data reconstruction errors. It performs experimental evaluation tests on clustering using multi-valued data sets. The results demonstrate that the proposed method has achieved high performance in clustering compared to three state-of-the-art feature selection algorithms. It shows better achievement than then the existing clustering algorithm. In the feature, this approach can be extended to assess cluster performance in an unsupervised feature selection requirements unconditionally.

REFERENCES

1. M. Almalawi, A. Fahad, Z. T., Muhammad A. Cheema, I. Khalil "k-NNVWC: An Efficient k-Nearest Neighbors Approach Based on Various-Widths Clustering," IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 1, Jan. 2016.
2. J. Wang, J. Wang, J. Song, X.-S. Xu, H. T. Shen, S. Li, "Optimized Cartesian K-Means," IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 1, Jan. 2015
3. H. Jaber, F. Marle, M. Jankovic, "Improving Collaborative Decision Making in New Product Development Projects Using Clustering Algorithms" IEEE Transactions On Engineering Management, Vol. 62, No. 4, Nov. 2015.
4. T.-Hien T. Nguyen, V.-N. Huynh, "A k-Means-Like Algorithm for Clustering Categorical Data Using an Information Theoretic-Based Dissimilarity Measure," Springer International Publishing Switzerland, 10.1007/978-3-319-30024-5, pp. 115-130, 2016.
5. Z. Li, Jing Liu, Yi Yang, X. Zhou, and H. Lu, "Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection," IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 9, Sept. 2014.
6. Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "L2, 1-norm regularized discriminative feature selection for unsupervised learning," In Proc. Int. Joint Conf. Artif. Intell., Vol. 22, p. 1589, 2011.
7. J. Wu, H. Liu, H. Xiong, J. Cao, J. Chen, "K-Means-Based Consensus Clustering: A Unified View," IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 1, Jan. 2015.
8. B. Jiang, J. Pei, Y. Tao, X. Lin, "Clustering Uncertain Data Based on Probability Distribution Similarity," IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 4, April 2013.
9. L. Chen, Q. Jiang, S. Wang, "Model-Based Method for Projective Clustering," IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 7, July 2012.
10. Natthakan, T. Boongoen, S. Garrett, C. Price, "A Link-Based Cluster Ensemble Approach for Categorical Data Clustering," IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3, March 2012.
11. F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint L2, 1-norms minimization," In Proc. Adv. Neural Inf. Process. Syst., vol. 23, pp. 1813-1821, 2010.
12. R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification," New York, NY, USA: Wiley, 2012.
13. O. M. San, Huynh, V.N., Nakamori Y., "An alternative extension of the k-means algorithm for clustering categorical data," Int. Journal Application Math. Computation Science. 14, 241- 247, 2004.

14. D. Ienco, R. G. Pensa, R. Meo, "From context to distance: learning dissimilarity for categorical data clustering," ACM Trans. Knowledge Discovery Data, 6(1),1-25, 2012.
15. J. Pei, B. Jiang, X. Lin, and Y. Yuan, "Probabilistic Skylines on Uncertain Data," Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB), 2007.
16. Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," In Proc. AAAI, 2010.
17. W. Fan, N. Bouguila, D. Ziou, "Unsupervised hybrid feature extraction selection for high-dimensional non-Gaussian data clustering with variation inference," IEEE Tran. Know. Data Eng.., Vol. 25, No. 7, pp. 1670-1685, Jul. 2013.
18. L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach," J. Mach. Learn. Res., vol. 6, pp. 1855-1887, 2005.
19. Y. Tao, R. Cheng, X. Xiao, W.K. Ngai, B. Kao, and S. Prabhakar, "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2005
20. D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," In Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 333-342, 2010.
21. D. Tao, X. Li, X. Wu, and S. J. Maybank, "General averaged divergence analysis," In Proc. 7th IEEE Int. Conf. Data Mining, pp. 302-311, 2007.
22. X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," In Proc. Adv. Neural Inf. Process. Syst., vol. 186, p. 189, 2005.
23. S. Boutemedjet, D. Ziou, and N. Bouguila, "Unsupervised feature selection for the accurate recommendation of high-dimensional image data," in Proc. Advanced Neural Inf. Process. Syst., 2007.
24. J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," The Am. Statist., vol. 42, no. 1, pp. 59-66, 1988.
25. Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in Proc. 24th Int. Conf. Mach. Learn., pp. 1151-1157, 2007.
26. B. Krishnapuram, A. J. Hartemink, L. Carin, and M. A. T. Figueiredo, "A Bayesian approach to joint feature selection and classifier design," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 9, pp. 1105-1111, Sept. 2004.
27. X. Li and Y. Pang, "Deterministic column-based matrix decomposition," IEEE Trans. Knowl. Data Eng., vol. 22, no. 1, pp. 145-149, Jan. 2010.
28. W. Liu, D. Tao, and J. Liu, "Transductive component analysis", In Proc. 8th IEEE Int. Conf. Data Mining, pp. 433-442, 2008.
29. F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection", In Proc. AAAI, 2008, vol. 2, pp. 671-676, 2008.
30. X. He, M. Ji, C. Zhang, and H. Bao, "A variance minimization criterion to feature selection using Laplacian regularization", IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 10, pp. 2013-2025, Oct. 2011.
31. Y. Guan, M. I. Jordan, and J. G. Dy, "A unified probabilistic model for global and local unsupervised feature selection", in Proc. 28th Int. Conf. Mach. Learn., pp. 1073-1080, 2011.
32. J. Tang, X. Hu, H. Gao, and H. Liu, "Discriminant analysis for unsupervised feature selection", SDM, SIAM, pp. 938-946, 2014.
33. Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis", in Proc. SDM, 2007.
34. H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering", IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491-502, Apr. 2005.
35. A. Jain and D. Zongker, "Feature Selection: Evaluation, application, and small sample performance", IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 2, pp. 153-158, Feb. 1997.
36. A. E. Brodley and J. G. Dy, "Feature selection for unsupervised learning", International Journal Machine Learning Res., vol. 5, pp. 845-889, Dec. 2004.
37. R. Cheng, D.V. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2003.
38. S. B. Kotsiantis, P.E. Pintelas, "Recent Advances in Clustering: A Brief Survey", WSEAS Trans. Information Science and Applications, vol. 11, no. 1, pp. 73-81, 2004.

AUTHORS PROFILE

P Gopala Krishna is a Research Scholar in the Department of Computer Science and Systems Engineering, College of Engineering(Autonomous), Andhra University, pursuing Ph.d in the area of Data Mining

Dr D Lalitha Bhaskari is a Professor of CS & SE, Department of CS & SE, College of Engineering (Autonomous), Andhra University. Recipient of “Young Engineer Award” by the Institute of engineers (INDIA) in the year 2008 in the field of Computer Science during 2008. Her areas of research interest includes Data Security, Image Processing, Pattern Recognition, Digital Watermarking and Image mining.