Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 1 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Survey on Clustering Techniques for Multi- Valued Data Sets

By Lnc.Prakash K, K.Anuradha & D.Vasumathi

Annamacharya Institute of Technology and Sciences

Abstract- The complexity of the attributes in some particular domain is high when compare to the standard domain, the reason for this is its internal variation and the structure .their representation needs more complex data called multi-valued data which is introduced in this paper. Because of this reason it is needed to extend the data examination techniques (for example characterization, discrimination, association analysis, classification, clustering, outlier analysis, evaluation analysis) to multi-valued data so that we get more exact and consolidated multi-valued data sets. We say that multi-valued data analysis is an expansion of the standard data analysis techniques. The objects of multi-valued data sets are represented by multi-valued attributes and they contain more than one value for one entry in the data base. An example for this type of attribute is “languages known” .this attribute may contain more than one value for the corresponding objects because one person may be known more than one language.

GJCST-C Classification : H.3.3

A Survey on Clustering Techniques for Multi-Valued Data Sets

Lnc.Prakash K [α], K.Anuradha [σ] & D.Vasumathi [ρ]

Abstract- The complexity of the attributes in some particular domain is high when compare to the standard domain, the reason for this is its internal variation and the structure .their representation needs more complex data called multi-valued data which is introduced in this paper. Because of this reason it is needed to extend the data examination techniques (for example characterization, discrimination, association analysis, classification, clustering, outlier analysis, evaluation analysis) to multi-valued data so that we get more exact and consolidated multi-valued data sets. We say that multi-valued data analysis is an expansion of the standard data analysis techniques. The objects of multi-valued data sets are represented by multi-valued attributes and they contain more than one value for one entry in the data base. An example for this type of attribute is "languages known" .this attribute may contain more than one value for the corresponding objects because one person may be known more than one language.

I. Introduction

In the process of making more general surveys the persons are permitted to give more than one answer for a particular question. The answer may be a set of categorical values or a set of numerical values or sometimes it may be the combination of both. Now a days the data base is going on increasing and it is needed the integration of different data bases as a result a very lengthy databases are formed. The data bases that are formed in this way may contain the same objects that are repeated many times with different values .the increasing importance is to reduce the data base size by summarizing the data without loss of the information that is used for analysis. This can be achieved by introducing the concept of multi-valued attributes in the data bases. Because the cells of such data may contain not only single numerical or categorical values, but much more multifaceted information, such as subsets of categorical variable values, intervals of ordinal variable values, dependencies and need rules to be specified. These new statistical units are called multi-valued objects and there is a need for an extension of standard Data Analysis to such objects, called multi-valued Data analysis.

II. Input to Multi-Valued Data Analysis Algorithms

In general Columns of the source data table are variables (attributes) and rows (objects) are multi-valued descriptions. Each cell of this multi-valued data table may contain data of different types as given bellow.

- The sell may contain a quantitative value for example. If "height" is an attribute and a is a unit then height (a)=2.4 .
- The sell may contain discrete value, for instance, consider "village" is a variable and a is a unit then village (a) = Raja pet.
- The sell may contain Multi-valued, for example, in the quantitative case: height (a) =(3 .2, 4 .1,5)which means that the weight of a maybe 3 .2 or 4 .1 or 5 . In the discrete case, language (a) = (Telugu, Hindi, Tamil) means that the language of a may be Telugu or Hindi or Tamil. Here it is clear that the above two are special cases of this case.
- The sell may contain Interval valued data: for instance height (a) = [12, 15], which tells that the height of a varies in the interval [12, 15].

Some other types of variables may also exist with Multi valued weights, Taxonomic, Hierarchically dependent, logically dependent.

III. Result of Multi-Valued Data Analysis Algorithms

The output of any data analysis depending on the type of the data mining task. If the task is characterization it represents the nature of data and its character. When the task is association we can find the frequent items and then form the association rules. If the task is classification we can find the different classes that are associated with the data. If the task is clustering we can partition the data into different similar groups depending on the similarity measure.

IV. Source of Multi-Valued Data

Multi-valued data is formed from the tables that are integrated from different sources so this type of data is summarized and concise so the requirement of consolidation of different huge data sets causes multi-valued data sets. The result from several probability distributions, percentiles and the range of any random variables also produces multi-valued data. Some times

Author α: Research scholar(JNTUH), Department of Computer Science and Engineering, AITS Rajampet, AP, India.
e-mail: klnc.prakash@gmail.com
Author σ: Professor, Department of Computer Science and Engineering,GRIET, Hyderabad, TS, India. e-mail: kodali.anuradha@yahool.com
Author ρ: Professor, Department of Computer Science and Engineering, JNTUCEH, Hyderabad,TS, India. e-mail: rochan44@gmail.com

Global Journal of Computer Science and Technology (C) Volume XVI Issue V Version I Year 2016

in order to answer to a query that is applied on different relational data bases form the multi-valued data sets. Multi-valued data can also be generated from Data Analysis (functional analysis, clustering, associations, neural networks . . .) of standard databases, from expert knowledge (scenario of traffic accidents, type of employed . . .), from time series (in the intervals of time), from private data (in the view of hiding the initial data by less precision), etc.

Table a

pan number	person's name	Age	category
11231	A.Ajay kumar	21.0	Male
14561	A.Vijay kumar	22.0	Male
17891	A.Arun kumar	23.0.	Male
11011	A.Varun kumar	24 .0	Male
11021	S.Swathi	25.0	Female
11031	S.Swetha	25.0	Female

Table b

pan number	City	Mode of payment	Date of payment	Amount in Rs
11231	Visakhapatnam(V)	Physical cash	1/06/13	500.00
11231	Hyderabad(H)	Credit to bank	10/07/13	80.00
11231	Nellore(N)	Check	11/07/13	100.00
14561	Visakhapatnam(V)	Physical cash	19/08/13	400.00
14561	Anakapalli(A)	Credit to bank	20/08/13	200.00
17891	Anakapalli(A)	Credit to bank	5/09/13	50.00
11011	Vizianagaram(Vz)	Physical cash	8/09/13	25.00

Table c

Name	Age	Category	Mode of payment	Amount in Rs	City
A.Ajay kumar	21.0	Male	Physical cash	500.00	Visakhapatanam(V)
A.Ajay kumar	21.0	Male	Credit to bank	80.00	Hyderabad(H)
A.Ajay kumar	21.0	Male	Check	100.00	Nellore(N)
A.Ajay kumar	22.0	Male	Physical cash	400.00	Visakhapatanam(V)
A.Ajay kumar	22.0	Male	Credit to bank	200.00	Anakapalli(A)
A.Arun kumar	23.0	Male	Credit	50.00	Anakapalli(A)
A.Varun kumar	24.0	Male	Physical cash	25.00	Vizianagaram(Vz)
A.Varun kumar	14.0	Male	Credit to bank	50.00	Visakhapatanam(V)

There are some approaches like ribeiro95 [1] and Thompson 91[5] identify the knowledge directly from domains that are structured. The most straight forward way to form a linear file is joining of related tables. the above table-c is the result of both of the tables table-a and table-b. Table-c is obtained by applying the natural join to both of the tables table-a and table-b which has the problem that a single object cannot be identified unique, that means a single object can be represented by more than one row. The problem with this type of representation is most of the data mining and machine learning algorithms identify each row as a unique object even though an object is identified by more than one tuple.

So, Therefore this construction between a structure database and linear file is understood by many analysis frame works in efficiently This paper introduces a data mining framework and a information discovery to deal with this drawback of the previous linear file representations and we finally concentrate on the issues of structured database and discovery of a set of queries that describe the features of objects in the database. To overcome from this problem mentioned above it is needed to develop a better illustration procedure which represents information for objects that are co-related with other objects.

One simple solution to solve the problem is to combine the associated objects into a single object by applying some aggregate operations. The issue in this concept is the selection of aggregate function that can be applied and minimizing the Loss of technical information because of the aggregate function so finally there forms a cluster of related objects that represent a set of objects as a single object. To solve the same problem another technique is introduction of multi-valued attributes for the objects. The generalization of this type of linear file arrangement is as multi-valued data set. in this types of data bases the values of multi-valued attribute are either a single value or a bin of values these bin of values represent the data that associate with the object with respect to the multi-valued attribute. Table-d consists the multi-valued data that is generated from the two tables table-a and table-b, in this table product type, purchase place and purchase amount are multi-valued attributes, the objects have a set of values in between braces in the corresponding cells of these multi-valued attributes, if there is no any value we represent it by null or if there is a single value we put that value as it is.

Table d: data set with multi-valued of attributes

Name	Age	Product type	category	Purchase place	Purchase amount
A.Ajay kumar	21.0	{physical cash,credit to bank,check}	Male	{Visakhapatnam(V),Hyderabad(H), Nellore(N)}	{500.00,80.00,100.00}
A.Vijay kumar	22.0	{ physical cash, credit to bank }	Male	{Visakhapatnam(V),anakapally(A)}	{400.00,200.00}
A.Arun kumar	23.0	{Credit credit to bank}	Male	Anakapally(A)	50.00
A.Varun kumar	24.0	{ physical cash, credit to bank }	Male	{vizianagaram(Vz),Visakhapatnam(V)}	{25.00,50.00}

Now table-d contains the unique tuples for representing the data of one person which is different from the data that is denoted by table-c. So a useful set of queries can be applied to this type of data for implementing data mining tasks, because of multi-valued attributes are existed in the data base the following problems are encountered during this process.

- It is needed to identify the set of queries that access the features of objects in the database.
- Applying the data mining techniques to the database that consists of multi-valued attributes.

Most of the data mining techniques that are existed cannot work properly on the multi-value datasets which consists of a set of values because they are developed for applying on single valued attributes. So the need is to develop the techniques that can be applied on multi valued attributes.

Multi-valued Datasets that are available in UCI machine learning repository are given in the bellow table.

Table e

Name of the Dataset	Number of classes	Number of normal attributes	Number of multi - valued attributes	Number of binary attributes
“Promoter”	106	2	57	-
“Hayes-Roth”	160	3	4	-
“Breast cancer”	286	2	7	3
“Monks-1”	432	2	4	2
“Monks-2”	432	2	4	2
“Monks-3”	432	2	4	2
“Balance”	625	3	4	-
“Soya large”	683	19	19	16
“Tic-tac-toe”	958	2	9	-
“Car”	1728	4	6	-
“DNA”	3190	3	60	-
“Mushroom”	8124	2	18	4
“Nursery”	12960	2	7	1

V. Similarity Measures for Multi-Valued Data Sets

In general the categorization of two types of attributes in the data bases are quantitative attributes and qualitative attributes to apply data mining tasks on these databases we are needed to find the similarity measures among these type of attributes .

VI. Qualitative Type

Tversky (1977) proposed a contrast model and ratio model that is implemented by generalizing a several set of theoretical similarity models .Tversky described the objects as sets of features as a replacement of geometric points in a geometric space. To demonstrate his models, let *a* and *b* are two objects, and *X* and *Y* indicate the sets of features associated with the objects *x* and y respectively. Tversky proposed the following family of proximity measures which is called the contrast model:

$$S(x, y) = \theta\ n\ (X \cap Y)\ -\alpha\ n(X - Y)\ -\ \beta\ n(X - Y)$$

For some $\theta, \alpha, \beta \geq 0$; *n* is usually the cardinality of the set. In the previous models, the proximity between objects was determined only by their common features, or only by their distinctive features. In the contrast model, the proximity of a pair of objects is represented as a linear combination of the measures of the common and the distinctive features. The contrast model denotes proximity between objects as a weighted difference of the measures for their common and distinctive features. The given bellow family of similarity measures denotes the ratio model:

$$S(x, y) = n\ (X \cap Y)\ /\ [n(X \cap Y) + \alpha\ n(X - Y) + \beta\ n(Y - X)],\ \alpha,\ \beta \geq 0$$

In the ratio model, the proximity value is normalized to a value range of *0* and *1*. In Tversky's set theoretic similarity models, a feature usually denotes a value of a binary attribute or a nominal attribute but it can be extended to interval or ordinal type. For the qualitative type of multi-valued case, Tversky's set proximity can be used since we can consider this case as an attribute for an object has group feature property (e.g., a set of feature values).

VII. Attributes That are Quantitative Type

To find the proximity within the group when the attributes are multi-valued type we use group mean for the particular attribute with respect to the object by using Euclidean distance like measures but the problem with this method is it should not consider the cardinality of the elements in a group. Another approach towards this is group average which can be used to calculate inter-group proximity. In this approach group similarity is calculated by taking the average of all the inter-object measures for those pairs of objects from which each object of a pair is in diverse groups.

For example, the average dissimilarity between group P and Q can be defined as given bellow

$$D\ (P,\ Q) = \textstyle\sum_{i=1}^{n} d(p,q)/n),$$

where *n* is the cardinality of object-pairs, *d* (*a*, *b*) is the variation metric for the i^{th} pair of objects p and q where p∊ *P* And q ∊ Q. In calculating group similarity using on group average, decision on whether we compute the average for every probable pair of similarity or the average for a subset of possible pairs of similarity is required.

VIII. Algorithms for Clustering

This paper is going to present different algorithms for clustering by considering the properties of multi-valued Data characteristics like formation, noise, dimensionality of the attributes, algorithm implementation, dimension of the data, and shape of the cluster. The overview of algorithms for clustering is given bellow.

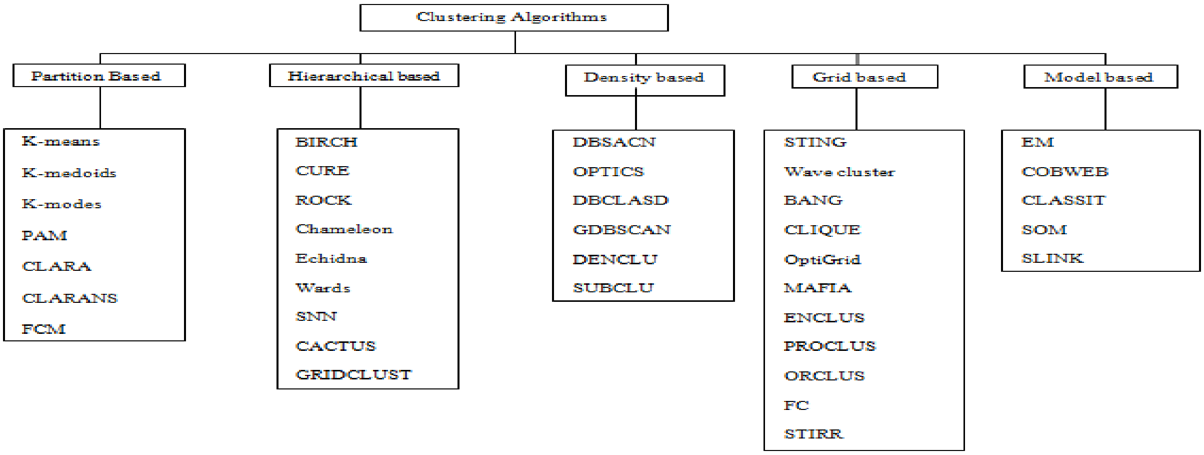

Figure 1

IX. Partitioning Algorithms for Clustering

The data objects in these types of methods are initially considered as a single cluster and this can be partitioned into different clusters depending on the number of clusters required. The objects are assigned into partitions by iteratively placing the points between the partitions. There are different re arrangement schemes that iteratively reassign the points among the pre defined number of clusters not like hierarchical methods clusters not allowed revisiting the previously formed clusters. These algorithms Progressively develop the quality of clusters.

In all of the partitioning algorithms the number of clusters is decided initially, if we want to find the number of clusters automatically we can think data comes from the mixture of various probability distributions .the major advantage of probabilistic approach is interpretability of the clusters that are formed. The summarized cluster representation can be done by allowing the measures of intra cluster similarity, an another approach to measure the quality of clustering is the definition of objective function . in general partitioning algorithms constructs the clusters in non-convex shape.

Some partitioning algorithms for clustering are K_means, K_modes, K_medoids (PAM, CLARA, CLARANS, and FCM). Partition based algorithms can found clusters of Non convex shapes. The pros and cons of partitioning clustering methods are described bellow.

Advantages:

- Comparatively scalable and simple.
- These algorithms are suitable for the datasets which forms spherical shaped clusters that are well-separated

Disadvantages:

- These algorithms causes efficiency degradation in the high dimensional spaces because almost all pairs of points are at a distance of its average the reason for this is the distance between points in high dimensional spaces is not defined properly.
- The description of cluster is less explained by these algorithms.
- User needs to specify the number of clusters in advance.
- More sensitive to initialization phase that means final results depends on the clusters that are formed at initial, also sensitive to noise and outliers.
- These are in efficient to deal with non-convex shaped clusters of unreliable size and density.

X. Hierarchical Algorithms for Clustering

Hierarchical clustering algorithms are of two types they are Agglomerative (top-bottom) and Divisive (bottom- top). In Agglomerative Hierarchical clustering initially one object is selected and the other objects are embedded consecutively to form bigger clusters the merging of objects depends on the distance like maximum, minimum and average .this procedure is repeated number of times until to form desired number of clusters. some Hierarchical algorithms for clustering are BIRCH, CURE, ROCK, Chameleon, Wards, SNN, GRIDCLUST, CACTUS in which clusters of Non convex shaped and, Arbitrary Hyper rectangular are produced.

Some times to find the proximity between sub sets it is needed to generalize the proximity between individual points. Such proximity measure is called linkage metric. This is also the case when the data base consists of multi-valued data. This type of metric impacts the Hierarchical clustering algorithms because it impacts the closeness and relation of connectivity among the attributes without considering its structure.

Mostly used inter-cluster linkage metrics are single link, average_ link, and complete link. The principal proximity of dissimilarity measure (usually distance) is computed for every pair of points with one point in the first set and another point in the second set. A specific functional operation such as minimum Average or maximum is applied to pair-wise proximity of dissimilarity measures an example for such functional operation is given bellow:

$$D\ (C_1,\ C_2\ )=\text{ functional operation }\{d\ (x,y)/\ x\in C_1,\ y\in C_1\}$$

The pros and cons of partitioning clustering methods are described bellow

Advantages:

- These algorithms are flexible with respect to the level of granularity.
- similarity or distance proximities are easily managed.
- These algorithms can be applicable to any types of attributes.

Disadvantages:

- It is difficult to identify the termination point.
- Most of the hierarchical algorithms cannot be repeated if once constructed.
- Clusters are formed depending on the reason of their improvement.

XI. Density based Algorithms for Clustering

In density based methods the set of points which are in the space of Euclidean are partitioned into a set of points depending on the density among the points, connectivity between each pair of points and the boundary of the points that are grouped. In density based clusters each point in the cluster closely related to its adjacent neighbor, the clusters are formed in any direction that the density of the points in one cluster should be maximum. Because of this reason the cluster that depends on density are of arbitrary shaped as a result it provides a good security against outliers.

These algorithms classify the data objects into core_points, border_points and noise points to form dense regions. The clusters are formed by connecting the core points together. So there may be a chance of forming non spherical or arbitrary shaped clusters. Some well known density based algorithms for clustering are "DBSCAN", "OPTICS","DBCLASD", "GDBSCAN" ," DENCLUE" and "SUBCLUE".

XII. Advantages of Density based Algorithms for Clustering (DBSCAN)

1. All of these types' algorithms like DBSCAN are not needed to specify the number of clusters in the data in advance, which is opposite to k-means.
2. The major advantage of DBSCAN like algorithms are it finds arbitrary shaped clusters. It can even form clusters completely enclosed by (but not connected to) another different cluster. DBSCAN can perform well even though there is noise in the data and is robust to outliers.
3. The queries can be easily and speedily processed on the clusters that are formed by "DBSCAN".

XIII. Disadvantages of Density based Algorithms Clustering (dbscan)

1. DBSCAN is not entirely deterministic: Depending on the order of the data processed different types of clusters are formed, sometimes border points may belongs to more than one cluster so it can be a part of either of the cluster.
2. The quality of DBSCAN is influenced by the distance metric used in the function for finding the neighborhood in the Query. The distance metric that is used to find the distance is Euclidean distance for multi-valued data and high dimensional data.
3. If the data objects have high variation in its density then DBSCAN cannot cluster that data sets well. Since the combination of minPts and neighborhood distance ε cannot be chosen appropriately for all clusters.

XIV. Grid based Algorithms for Clustering

The density based methods needs some clarifications like density, boundary and connectivity for grouping the data objects. Another way of representation is by the consideration of attribute space. To minimize the search combination and space many different rectangular segments are considered where each segment is the Cartesian of individual different sub spaces.

The methods that partition the space into different cells or segments are called grid-based methods. Here each segment is related to a single value or a group of values (if the data is multi-valued data) is considered as a unit, as a result the concentration is turned on to the partition space. The partition is based on the characteristics of the grid that is formed from the data. The advantage of this type of arrangement of data in the form of grids is it does not depend on the order of the points that are arranged in the form of grids.

When the data is of numeric type then only density based and partition based methods give results efficiently but grid based techniques perform well for the attributes with different data types along with multi-valued data. For better performance the density based method "DENCLUE" initially uses grid structure. To structure real clusters Grid based algorithms uses subspace and hierarchical techniques for clustering. When compare to all Clustering algorithms Grid algorithms are very speedy and efficient processing algorithms. Arbitrary shaped clusters are formed by Adaptive grid algorithms such as MAFIA and AMR by using the grid cells.

XV. Model based Clustering Algorithms

A model is hypothesized for each of the clusters and tries to find Set of data points are related together based on different strategies like conceptual methods and statistical methods. For model based algorithms the well known approaches are one is neural network approach and another one is statistical approach. Algorithms such as "EM", "CLASSIT", "COBWEB SOM", and "SLINK" are some Model based clustering algorithms.

XVI. Research Challenges

We have previously discussed that the problem of clustering becomes very demanding,attractive and at the same time challenging, when the data is of type categorical attributes and multi-valued attributed. The number of algorithms for the discovery of groups in such data is restricted, confined and limited, compared to the research devoted on data sets with numerical data. Further, few algorithms deal with mixtures of values, i.e., attributes of numerical and categorical values. Ideally, a clustering algorithm should be needed to

- scale well, i.e., at most one scan of the data is necessary;
- handle deviations professionally;
- discover arbitrary-shaped clusters;
- give sensible execution times in the attendance of high dimensional data sets;
- present a concise model of the clusters;

XVII. Summary and Conclusion

In this paper, we analyzed the difficulty of generating single flat file arrangement to represent Data sets that have been generated from structured databases, and pointed out its inappropriate representation to represent related information, a fact that has been frequently overlooked by recent data mining investigation. To overcome these difficulties, we used a better representation scheme, called multi-valued *data set*, which allows attributes of an object to have a *set* of values, and studied how existing similarity measures for single-valued attributes could be applied to measure group similarity for multi-valued *data sets* in clustering.We also proposed a unified framework for proximity measures to work with multi-valued *data sets* with mixed types. Once the target database is grouped into clusters with similar properties, the discriminate query detection system, MASSON can find out useful characteristic information for a set of objects that fit in to a cluster. We claim that the planned representation scheme is suitable to cope with related information and that it is more communicative than the traditional single flat file format.

More prominently, the relationship information in a structured database is actually considered in clustering procedure.

References Références Referencias

1. Ribeiro, J.S., Kaufmann, K., and Kerschberg, L. (1995). Knowledge Discovery from Multiple Data bases, In Proc. Of the 1st Int'l Conf. On Knowledge Discovery and Data Mining, Quebec, Montreal.
2. Ryu, T.W and Eick, C.F. (1996a). Deriving Queries from Results using Genetic Programming, In Proceedings of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining. Portland, Oregon.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.
3. Ryu, T.W and Eick, C.F. (1996b). MASSON: Discovering Commonalities in Collection of Objects using Genetic Programming, In Proceedings of the Genetic Programming 1996 Conference, Stanford University, San FranciscoR. Cheng, D.V. Kalashnikov, and S. Prabhakar, X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection", In Proc. Adv. Neural Inf. Process. Syst., vol. 186, p. 189, 2005.
4. Ryu ,T.W. and Eick ,C.F. (1998). Automated Discovery of Discriminate Rules for a Group of Objects in Databases, In Conference on Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA, June 11-13.R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", New York, NY, USA: Wiley, 2012.
5. Thompson, K., and Langley, P. (1991). "Concept formation in structured domains, In Concept Formation: Knowledge and Experience in Unsupervised Learning", Eds., Fisher, D.H; Pazzani, M.; and Langley, P., Morgan Kaufmann.
6. Tversky, A. (1977). "Features of similarity, Psychological review", 84(4): 327-352, July.
7. Everitt, B.S. (1993). "Cluster Analysis, Edward Arnold, Copublished by Halsted Press and Imprint", of John Wiley & Sons Inc., 3rd edition.
8. Gower, J.C. (1971). "A general coefficient of similarity and some of its properties", Biometrics 27, 857-872.
9. Koza, John R. (1990). "Genetic Programming: On the Programming of Computers by Means of Natural Selection", Cambridge, MA: The MIT Press.
10. S.B. Kotsiantis and P.E. Pintelas, "Recent Advances in Clustering: A Brief Survey",WSEAS Trans. Information Science and Applications, vol. 11, no. 1, pp. 73-81, 2004.
11. 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.
12. 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.
13. 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.
14. 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.
15. 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.
16. X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection", In Proc. Adv. Neural Inf. Process. Syst., vol. 186, p. 189, 2005.
17. 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.
18. R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", New York, NY, USA: Wiley, 2012.
19. X. Li and Y. Pang, "Deterministic column-based matrix decomposition", IEEE Trans. Knowl. Data Eng., vol. 22, no. 1, pp. 145-149, Jan. 2010.
20. W. Liu, D. Tao, and J. Liu, "Transductive component analysis", In Proc. 8th IEEE Int. Conf. Data Mining, pp. 433-442, 2008.

This page is intentionally left blank