International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.4 (2016), pp.81-90
http://dx.doi.org/10.14257/ijsip.2016.9.4.08

A Novel Multilayer Data Clustering Framework based on Feature Selection and Modified K-Means Algorithm

Ganglong Duan, Wenxiu Hu and Zhiguang Zhang

Xi'an University of Technology, Shaanxi 710054, China
duanganglong1@126.com

Abstract

With the rapid development of computer science and technology, the data analysis technique has been a hottest research area in the pattern recognition research community. Cluster analysis is an important step in data mining. For clustering, various multi-objective techniques are evolved, which can automatically partition the data. In this paper, we propose a novel multilayer data clustering framework based on feature selection and modified K-Means algorithm. To facilitate the clustering, the proposed algorithm selects a representative feature subset to reduce the dimension of the raw data set. Besides, the selected feature subset has fewer missing values than the raw data set, which may improve the cluster accuracy. Another unique property of the proposed algorithm is the use of partial distance strategy. The experimental analysis and simulation indicate the feasibility and robustness of our method, in the future, we plan to conduct more mathematical analysis to modify our algorithm to achieve better result.

Keywords*: Data Clustering, Feature Selection, K-Means Algorithm, Data Mining*

1. Introduction

Clustering algorithms have emerged and rapidly developed as an alternative powerful meta-learning tool to accurately analyze the massive volume of data generated by modern applications such as data analysis and mining [1-2], image classification [3-5] and computer vision [6]. In particular, their main goal is to categorize data into clusters such that objects are grouped in the same cluster when they are similar according to specific metrics [7]. Clustering algorithms can be divided into two categories, the traditional hard clustering and fuzzy clustering algorithm. Fuzzy clustering methods are popularly adopted due to their effectiveness and accuracy, the essential advantage is the expression ability and usage of optimization tools. For this purpose, and to validate the obtained partitioning, some cluster validity indices have been discussed in the literature. The validity of clusters should be such that it will be able to enforce an ordering of the clusters in terms of their goodness. In [8], Anuradha proposed a novel multi-objective framework for fuzzy clustering to achieve better classification accuracy. To achieve the goal, the K-means algorithm is adopted to initializing the initial data sets and clusters. These clusters are then optimized by using three objective functions as a fitness function in the traditional algorithm. Three objective functions such as compactness, connectedness, and symmetry of the cluster are optimized simultaneously. Their result indicates better performance compared with other state-of-the-art frameworks. In [9], Hu's group proposed a novel unsupervised possibilistic C-Means clustering to improve the efficiency of possibilistic c-means clustering (PCM) algorithm. A novel and robust clustering algorithm named as the weighted possibilistic c-means clustering (WPCC) algorithm is proposed to estimate the positions of centers of PCM accurately to serve for the coming clustering process. In [10], Chandrasekar proposed a two stage algorithm for data clustering. In their methodology, heuristic search algorithm is adopted to overcome and solve the problem of the black hole optimization. After reviewing more related literatures

ISSN: 2005-4254 IJSIP

in [11-23], we conclude and classify the current popular clustering algorithm in the figure 1.

In this paper, we propose a novel multilayer data clustering framework based on feature selection and modified K-Means algorithm. The rest of the paper is structured as the follows: section 2 gives the basic introduction to the clustering algorithm. Section 3 discusses our proposed novel algorithm and the in the fourth section, we conducts experimental analysis and simulation. In the final part, we conclude our methodology.

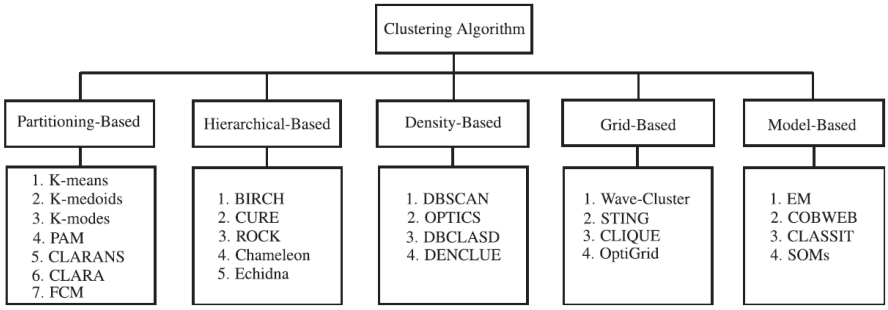

Figure 1.The Classification of Current Data Clustering Algorithms

2. Overview of Clustering Algorithms

In this section, we introduce three state-of-the-art clustering algorithms with detailed discussion and pseudo-code. The selected methods are: FCM, DENCLUE and EM algorithms which serve as the preliminary of our framework.

2.1. Fuzzy C-Means Algorithm

FCM [24] is a representative algorithm of fuzzy clustering which is based on K-means concepts to partition dataset into clusters. The FCM algorithm is a "soft" clustering method in which the objects are assigned to the clusters with a degree of belief. Therefore, an object can belong to more than one cluster with different degrees of belief. It tries to find the feature points in each cluster, named as the center of a cluster, then calculating the membership of each object in the cluster. Fuzzy c-means algorithm aims at minimizing the intra-cluster variance. It inherited problem of k - means, however, it is a local minimum and finally a cluster depends on the choice of initial weights. The FCM clustering is obtained by minimizing the objective function shown in the formula 1:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{m} \left| p_i - v_k \right|^2 \quad (1)$$

Where J denotes the objective function, n is the number of objects, c is the number of clusters, μ_{ik}^{m} is the likelihood values. The pseudo-code for FCM is shown in the table 1.

Table 1.The FCM Algorithm

Algorithm 1. Fuzzy C-means Algorithm (FCM)
1. **Input**: Given the dataset desire number of clusters, fuzzy parameters and stopping condition.
2. Calculate the cluster centroids and the objective value.
3. Compute the membership values stored in the matrix.
4. If the objective value of between consecutive iterations is less than the stopping condition, then *stop = true*.
5. **While** (!*stop*).
6. **Output**: A list of c cluster centers and a partition matrix are produced.

2.2. The DENCLUE Algorithm

The DENCLUE algorithm [25] analytically models the cluster distribution according to the sum of influence functions of all of the data points. The influence function can be seen as a function that describes the impact of a data point within its neighborhood. In this algorithm, clusters of arbitrary shape can be easily described a simple equation of kernel density function. The table 2 shows the DENCLUE algorithm.

Table 2. The DENCLUE Algorithm

Algorithm 2. The DENCLUE Algorithm
1. **Input**: The dataset, threshold T, the maximum diameter and the branching factor.
2. An initial in-memory CF-tree is constructed with one scan of the data.
3. Rebuild the CF-tree with a larger T.
4. Use the existing clustering algorithm on CF leaves.
5. Do additional passes over the dataset and reassign data points to the closest centroid.
6. **Output**: Compute CF points and the cluster result.

2.3. The Expectation-Maximization (EM) Algorithm

EM algorithm [26] is designed to estimate the maximum likelihood parameters of a statistical model in many situations, such as the one where the equations cannot be solved directly. The detailed steps are shown in the table 3.

Table 3.The Expectation-Maximization (EM) Algorithm

Algorithm 3. The Expectation-Maximization (EM) Algorithm
1. **Input**: The dataset (x), the total number of clusters (M), the accepted error for convergence (e) and the maximum number of iterations.
2. Compute the expectation of the complete data log-likelihood. $Q\left(\theta,\theta^T\right)=E\left[\log p\left(x^g,x^m\mid\theta\right)x^g,\theta^T\right]$
3. Select a new parameter estimate that maximizes the Q-function. $\theta^{T+1}=\arg\max_{\theta} Q\left(\theta,\theta^T\right)$
4. Increment t=t+1; repeat steps 2 and 3 until the convergence condition.
5. **Output**: A series of parameters represents the achievement of the convergence criterion.

3. Our Proposed Framework

3.1. Feature Selection Through Hierarchical Clustering

Irrelevant features, along with redundant features, severely corrupt the efficiency and the accuracy of data analysis and mining. In this section, the paper proposes a feature subset selection algorithm based on improved hierarchical clustering denoted as the FSHC. FSHC promotes cluster by eliminating the irrelevant and redundant features as much as possible. In addition, the selected feature subsets are missing value is less than the original data sets, it could improve the accuracy of the clustering algorithm. We firstly divides the raw dataset denote as the O into two sub-sets: C, I . Where $C = \{c_1, c_2, ..., c_t\}$ and $I = \{i_1, i_2, ..., i_t\}$. In the subset C, there are no objects with missing feature values, while each object has one or more missing feature values in the subset I.

The FSHC algorithm is generated based on two crucial steps. The first step is cluster features of the subset C into groups using the improved hierarchical clustering and the other one is to select representative features to final feature subset. In our methodology, we adopt the symmetric uncertainty (SU) as the measure of correlation between two features to cluster the features into individual groups. The definition of symmetric uncertainty (SU) is:

$$SU(X,Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)} \tag{2}$$

In which, $H(X)$ represents the entropy of a discrete random variable X . We introduce $p(x)$ to be the prior possibility for X . Hence, $H(X)$ is defined in the equation 3.

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{3}$$

In the formula 4, $Gain(X|Y)$ represents the amount by which the entropy of Y decreases is defined.

$$Gain(X|Y) = H(X) - H(X|Y) = H(Y) - H(X|Y) \tag{4}$$

Where, $H(X|Y)$ denotes the conditional entropy, and the formula 5 defines it.

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \tag{5}$$

FSHC uses bottom-up hierarchical clustering to cluster features into groups. Firstly, it will be the two most similar features merged into a cluster. The similarity between the two characteristics is measured using formula two. Then the recursive algorithm combined the two most similar clusters at each step. In the equation 6, we calculate the similarity between c_i, c_j , respectively under the prior assumption that there are m features in the cluster c_i and n features in the cluster c_j .

$$S_{ij} = \frac{1}{m \times n} \sum_{x \in c_i} \sum_{y \in c_j} |SU(x,y)| \tag{6}$$

Next, The FSHC algorithm selects one representative feature from every cluster to generate the final feature subset. It identifies the representative feature of each cluster by *F-Completeness*. We define the *F-Completeness* in the equation 7.

$$FC = 1 - m / n \tag{7}$$

The detailed steps of our proposed improved hierarchical clustering algorithm are concluded in the table 4.

Table 4.The Improved Hierarchical Clustering Algorithm

Algorithm 4. The Improved Hierarchical Clustering Algorithm
1. **Input**: The row dataset O , the clustering threshold θ .
2. Calculate similarity between every pair of features and merge the two most similar features into one.
3. **Repeat:**
i. Calculate similarity between every two clusters.
ii. Merge the two most similar into one cluster.
4. **Until:**
The maximum similarity is less than θ .
5. Calculate F-Completeness of every feature and set S to be Φ .
6. Add the features into S.
7. **Output**: Selected feature subset S.

3.2. Feature Selection Through Hierarchical Clustering

With the primary goal of enhancing clustering accuracy, we propose a novel k-means algorithm based on partial distance (NK-means algorithm). To measure the distance between the object O_k and the cluster center v_i , we define a novel distance metric paradigm denoted as the partial distance (PD) in the formula 8.

(8)

The parameter updating mechanism is defined in the equation 9 and the detailed steps of our NK-means algorithm are shown in the table 5.

$$v_{ij} = \sum_{O_k} I_k \times O_{kj} / \sum_{O_k} I_k$$
$$I_k = \begin{cases} 0, & if \quad o_{kj} = * \\ 1, & if \quad o_{kj} \neq * \end{cases} \qquad for \quad 1 \le j \le m, 1 \le k \le n \tag{9}$$

Table 5.The NK-Means Algorithm

Algorithm 5. The NK-means Algorithm
1. **Input**: The dataset O', the cluster number k.
2. Select randomly k objects as the initial cluster centers.
3. **Repeat:**
i. Calculate the distance between every object and each cluster center using 8.
ii. Distribute every object into its nearest cluster.
iii. Update all the features using 9.
4. **Until:**
The cluster centers keep constant.

5. **Output**: V-dataset consisting of cluster clusters;
K-dataset containing the labels of all the objects.

4. Experimental Analysis and Simulation

In order to verify the effectiveness and feasibility of our proposed methodology, we conduct numerical and experimental simulation in this section. We first artificially create missing values in the data sets for simulating incomplete data sets and then cluster them using the proposed algorithm. We take the accuracy and execution time into consideration to test the algorithm.

4.1. Set-up of the Experiment

The simulation environment is initialized as the follows. Six physical machines equipped with 2 TB hard disk and 8 GB of RAM, and the simulation software is installed on Windows Win8 platform and Intel core 4 quad core 3.2 GHz and 4 GB of RAM. To compare the advantages of the candidate clustering algorithms, eight simulated datasets are used in the experiments including: Multi-hop Outdoor Real Data (MHORD), Multi-hop Indoor Real Data (MHIRD), Single-hop Outdoor Real Data (SHORD), Single-hop Indoor Real Data (SHIRD). At the same time, three important measurement standards are adopted including: Cluster Accuracy (CA), Adjusted Rand Index (ARI) and Rand Index (RI).

4.2. Accuracy Experiment

The numerical result is shown in the table 6 and the visualized graph for the test result is shown in the figure 2.

Table 6. Quantitative and Numerical Evaluation For Our Algorithm

	CA				ARI				RI			
No	Ours	DEN	FCM	EM	Ours	DEN	FCM	EM	Ours	DEN	FCM	EM
1	**82.13**	66.13	71.78	79.85	**77.91**	58.85	62.63	72.02	**82.97**	77.15	79.36	81.76
2	**81.72**	66.43	70.67	80.13	**73.79**	55.36	60.79	71.44	**88.75**	76.29	81.11	84.07
3	**79.85**	63.29	77.66	78.26	**72.77**	53.01	66.25	70.33	**83.98**	76.56	80.29	81.14
4	80.12	67.53	73.73	**82.33**	**69.63**	59.37	63.10	66.98	**80.11**	76.57	73.29	79.26
5	**87.26**	64.66	79.77	77.01	**78.99**	61.23	62.15	76.31	**81.11**	74.46	76.13	80.08
6	**86.73**	64.45	70.76	79.73	**74.71**	50.93	62.97	73.07	**80.66**	79.29	76.12	75.31
7	**81.18**	68.98	74.10	80.92	**71.76**	55.47	62.60	70.56	**84.62**	75.50	81.80	79.73
8	**88.33**	64.62	77.36	85.98	**76.23**	53.10	70.21	75.23	**86.51**	81.21	82.33	84.04
9	**88.92**	66.86	77.45	87.22	**79.46**	51.21	70.15	77.12	**88.51**	86.22	80.11	85.68
10	**83.94**	69.94	74.31	80.69	**72.92**	50.35	69.07	71.81	**87.76**	80.22	86.13	82.75

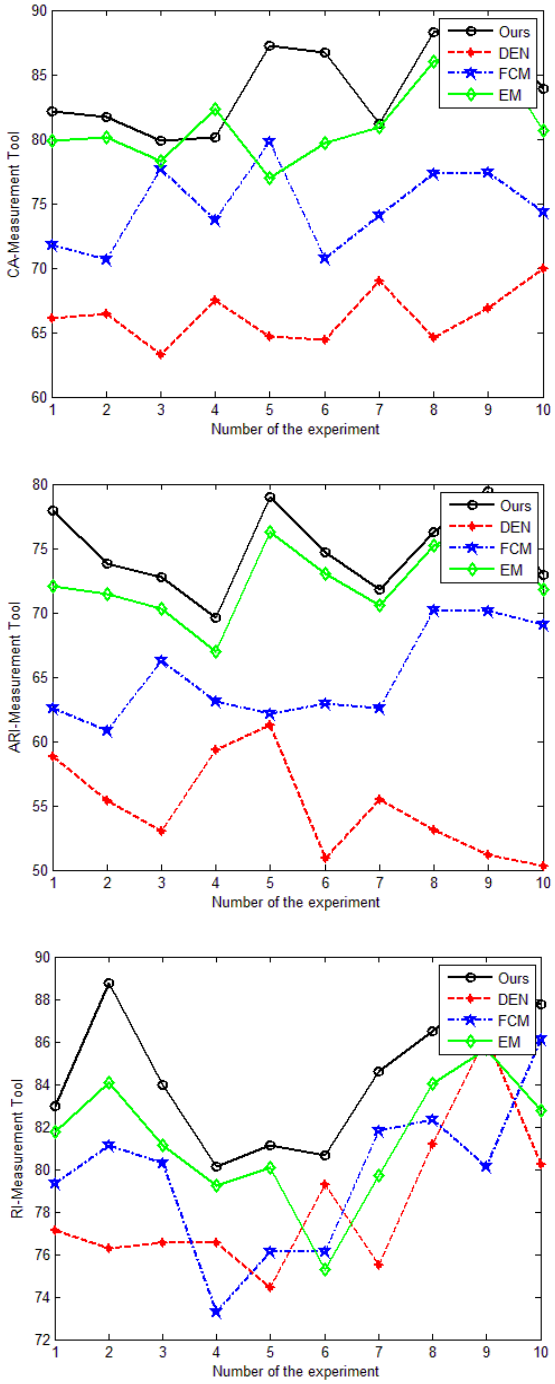

Figure 2.The Visualized Graph for the Experiment

4.3. Experimental Analysis on Execution Time

In this subsection, we present the execution time evaluate the efficiency of the proposed algorithm. The average execution time of the four algorithms is shown in the figure 3. We could derive that our algorithm is more efficient.

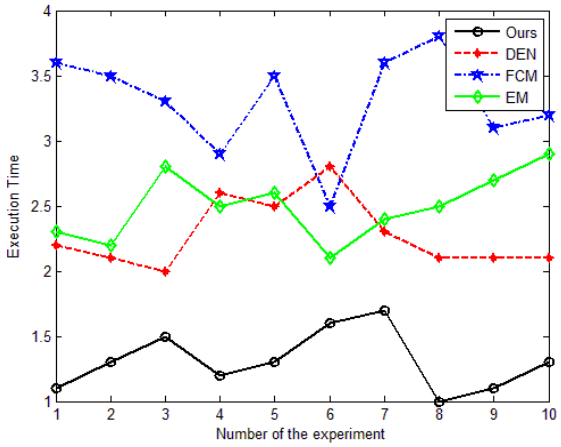

Figure 3.The Execution Time of the Four Algorithms

5. Conclusion and Summary

Clustering algorithms have emerged and rapidly developed as an alternative powerful meta-learning tool to accurately analyze the massive volume of data generated by modern applications. In this paper, we propose a novel multilayer data clustering framework based on feature selection and modified K-Means algorithm. To facilitate the clustering, the proposed algorithm selects a representative feature subset to reduce the dimension of the raw data set. Besides, the selected feature subset has fewer missing values than the raw data set, which may improve the cluster accuracy. Another unique property of the proposed algorithm is the use of partial distance strategy. The numerical experiment indicates that our proposed method outperforms other state-pf-the-art algorithms such as FCM and EM. Moreover, the execution time test shows that our method is not only robust but efficient as well. In the future, we plan to do more mathematical analysis for the updating procedure of clustering to achieve higher clustering accuracy.

Acknowledgements

This research is supported by the youth fund of national natural science fund project: information technology based on the dynamic adjustment speed economic value measurement and validation studies (NO. 71403206) and the project supported by the natural science foundation of the Shaanxi province education department: flexible recommendation system based on hybrid strategy research and development (NO. 14JK1522).

References

[1] D. T. Larose, “Discovering knowledge in data: an introduction to data mining”, John Wiley & Sons, **(2014)**.
[2] X.Wu, X. Zhu, G.-Q. Wu and W. Ding, "Data mining with big data," Knowledge and Data Engineering’, IEEE Transactions, vol. 26, no. 1, **(2014)**, pp. 97-107.
[3] X. Huang and Q. Lu, "A novel relearning approach for remote sensing image classification post-processing", In Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International, **(2014)**, pp. 3554-3557.
[4] H. Wang and J. Wang, “An effective image representation method using kernel classification,” in Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on November **(2014)**, pp. 853–858.
[5] R. Ji, Y. Gao, R. Hong, Q. Liu, D. Tao and X. Li, "Spectral-spatial constraint hyperspectral image classification." Geoscience and Remote Sensing, IEEE Transactions, vol. 52, no. 3, **(2014)**, pp. 1811-1824.

[6] H.Rositi, C. Frindel, M. Wiart, M. Langer, C. Olivier, F. Peyrin, and D. Rousseau, "Computer vision tools to optimize reconstruction parameters in x-ray in-line phase tomography", Physics in medicine and biology, vol. 59, no. 24, **(2014)**, pp. 7767-7775.
[7] A. Fahad, N. Alshatri, Z.Tari, A. ALAmri, A. Y Zomaya, I. Khalil, S.Foufou and A. Bouras, "A Survey of Clustering Algorithms for Big Data", Taxonomy & Empirical Analysis, **(2014)**.
[8] A. D. Thakare and M. A. More, "An improved data clustering algorithm in a multiobjective framework", In India Conference (INDICON), 2014 Annual IEEE, **(2014)**, pp. 1-5.
[9] J. Zhang and Ling Shen, "An Improved Fuzzy c-Means Clustering Algorithm Based on Shadowed Sets and PSO", Computational intelligence and neuroscience, **(2014)**.
[10] P. Chandrasekar and M. Krishnamoorthi, "BHOHS: A Two Stage Novel Algorithm for Data Clustering", In Intelligent Computing Applications (ICICA), 2014 International Conference on, IEEE, **(2014),** pp. 138-142.
[11] G. Krishnasamy, A. J. Kulkarni and R. Paramesran, "A hybrid approach for data clustering based on modified cohort intelligence and K-means", Expert Systems with Applications, vol. 41, no. 13, **(2014)**, pp. 6009-6016.
[12] P.-L. Lin, P.-W. Huang, C.-H. Kuo and Y. H. Lai, "A size-insensitive integrity-based fuzzy c-means method for data clustering", Pattern Recognition, vol. 47, no. 5, **(2014)**, pp. 2042-2056.
[13] H. Qin, X. Ma, T.Herawan and J. M. Zain, "MGR: An information theory based hierarchical divisive clustering algorithm for categorical data", Knowledge-Based Systems, vol. 67, **(2014)**, pp. 401-411.
[14] M.Yuwono, S.W. Su, B. D. Moulton and H. T. Nguyen, "Data clustering using variants of rapid centroid estimation." Evolutionary Computation, IEEE Transactions, vol. 18, no. 3, **(2014)**, pp. 366-377.
[15] R. Maronna, C. C. Aggarwal and C. K. Reddy (eds.), "Data clustering: algorithms and applications" Statistical Papers, **(2015)**, pp. 1-2.
[16] G. Q. Guo, W. J. Xiao and B. Lu, "Similarity Metric Based on Resistance Distance and its Applications to Data Clustering", In Applied Mechanics and Materials, vol. 556, **(2014)**, pp. 3654-3657.
[17] K. Adhikary, S. Das and S. Roy, "A Novel and Efficient Rough Set Based Clustering Technique for Gene Expression Data", In Business and Information Management (ICBIM), 2014 2nd International Conference on, IEEE, **(2014)**, pp. 41-46.
[18] X. Chen and C. Jian, "Gene expression data clustering based on graph regularized subspace segmentation", Neurocomputing, vol. 143, **(2014)**, pp. 44-50.
[19] P. A. Traganitis, K. Slavakis and G. B. Giannakis, "Big Data Clustering via Sketching and Validation" **,(2014)**.
[20] Y. Yang and Jianmin Jiang, "HMM-based hybrid meta-clustering ensemble for temporal data", Knowledge-Based Systems, vol. 56, **(2014)**, pp. 299-310.
[21] J. Zhang and L. Shen, "An Improved Fuzzy c-Means Clustering Algorithm Based on Shadowed Sets and PSO", Computational intelligence and neuroscience, **(2014)**.
[22] S.-H. Chang and C.-Y. Fan, "Analyzing Offshore Wind Power Patent Portfolios by Using Data Clustering." Industrial Engineeering & Management Systems, vol. 13, no. 1, **(2014)**, pp. 107-115.
[23] Z. M. Nopiah, A. K.Junoh, and A. K. Ariffin, "Vehicle interior noise and vibration level assessment through the data clustering and hybrid classification model", Applied Acoustics, vol. 87, **(2015)**, pp. 9-22.
[24] F. C. M.Sillé, L. Conde, J. Zhang, N. K. Akers, S. Sanchez, J. Maltbaek, J. E. Riby, M. T. Smith and C. F. Skibola, "Follicular lymphoma-protective HLA class II variants correlate with increased HLA-DQB1 protein expression", Genes and immunity, vol. 15, no. 2, **(2014)**, pp. 133-136.
[25] A. Hinneburg and D. A. Keim, "An ef cient approach to clustering in large multimedia databases with noise", in Proc. ACM SIGKDD Conf. Knowl. Discovery Ad Data Mining (KDD), **(1998)**, pp. 58-65.
[26] T. Denoeux, "Likelihood-based belief function: justification and some extensions to low-quality data." International Journal of Approximate Reasoning, vol. 55, no. 7, **(2014)**, pp. 1535-1547.

