2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering

Learning Scaling Coefficient in Possibilistic Latent Variable Algorithm from Complex Diagnosis Data

Zong-Xian Yin
Department of Multimedia and Entertainment Science
Southern Taiwan University
Tainan, Taiwan
yinzx@mail.stut.edu.tw

***Abstract*—The *Possibilistic Latent Variable (PLV) clustering algorithm* is a powerful tool for the analysis of complex datasets due to its robustness toward data distributions of different types and its ability to accurately identify the inherent clusters within the data. The scaling coefficient in the PLV algorithm plays a key role in reducing the effects of noise, thereby improving the precision of the clustering results. However, the optimal value of the scaling parameter varies depending on the population type of dataset. Accordingly, the current study proposes an evaluation method for evaluating suitable values of the scaling parameter. The relative comparison of each method is then examined by conducting PLV clustering trials using datasets comprising data of different types and patterns.**

Keywords-component; bioinformatics, clustering, machine learning, latent variable

I. INTRODUCTION

In an effort to keep pace with the increasing volume of biological and medical data made available by modern science, many researchers have considered the problem of how best to organize this information such that the end-users can locate the information they require in an efficient and reliable manner. Medical practitioners and researchers examine biological and clinical data for a variety of purposes, including detecting abnormal symptoms from the results of clinical trials, establishing relationships among gene expressions, constructing functional pathways from biological phenomena, and so forth. Many methods have been proposed for supporting users in performing these tasks. Of these methods, clustering is a particularly powerful technique for extracting meaningful insights from large volumes of raw data without the need for prior knowledge regarding the basic patterns within the data. Many clustering algorithms have been proposed for the analysis of biological and medical data.

Traditional clustering algorithms, such as k-Nearest Neighbor (k-NN) algorithm, Hierarchical Agglomerative Clustering (HAC), and so forth, are designed to assign each object within the dataset to a single cluster. By contrast, in fuzzy clustering techniques, each data point is assigned a fuzzy membership degree indicating its degree of belonging to each cluster within the dataset. Fuzzy clustering has found widespread application in a variety of fields nowadays, most notably in the computer vision and pattern recognition field [2-9].

The *Possibilistic Latent Variables* (*PLV*) clustering algorithm presented by the current authors in [1] combines a statistical estimation approach with the fuzzy degree of membership concept. PLV represents an ideal tool for the clustering of data acquired from practical applications due to its rapid convergence properties and its robustness toward the underlying distribution of the data. The performance of the PLV scheme is fundamentally dependent upon the value assigned to its scaling coefficient, β. In practice, an appropriate value of β prevents the centroids within the data from being skewed by "bad" samples and therefore improves the accuracy of the classification results. However, the optimal value of β depends on the distribution of the data within the dataset, and this is commonly unknown in advance. Accordingly, the current study presents a new method for evaluating an appropriate value of β such that the classification performance of the PLV scheme is enhanced. The respective performance of each method is examined by conducting clustering trials using synthetic and real-world medical / biological datasets comprising data of different types and distributions.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of PLV clustering algorithm. Section 3 describes the Annealing mechanism for evaluating the optimal value of its scaling parameter, β. Section 4 presents the clustering results obtained using the PLV algorithm with the mechanism. Finally, Section 5 presents some conclusions.

II. POSSIBILISTIC LATENT VARIABLE CLUSTERING ALGORITHM

The PLV clustering algorithm is based upon two basic hypothetical assumptions, namely (1) in clustering real-world datasets, the objects are not easily assigned to single clusters; and (2) the clusters are not necessarily Gaussian mixtures. Therefore, the assumptions can be made as:

$$\mu_{\cdot j} = \left\{ \mu_{ij} \middle| \mu_{ij} \in [0,1] \right\}^{c}, \text{ where } \sum_{i=1}^{c} \mu_{ij} = 1, i = 1,2,\cdots,c \quad (1)$$

978-0-7695-3656-9/09 $25.00 © 2009 IEEE
DOI 10.1109/BIBE.2009.61

where μ_{ij} represents the fuzzy degree of membership of object x_j to any given cluster C_i. The objective function applied in the clustering process can be formulated as

$$J(C,U;X)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\log\pi_i+\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\log Pos(x_j;C_i) \tag{2}$$

where m is the fuzziness of μ_{ij}. The possibilistic function $Pos(x_j;C_i)$ is the cost (or energy) of associating point x_j with cluster C_i. Thus, for all possible associations, the possibilistic function is given by

$$Pos(x_j;C_i)=\varpi_j\times\exp\left[-\beta\times d^2(x_j,c_i)\right]$$

$$\text{where } \varpi_j=\left[\sum_{k=1}^{c}\exp\left[-\beta\times d^2(x_j,c_k)\right]\right]^{-1}$$

where $d^2(x_j,c_i)=\left|c_i-x_j\right|^2$ is the square of the distance between x_j and the cluster center c_i. The remedial coefficient of $Pos(x_j;\mathbf{C})$ with respect to x_j is expressed by the weighted parameter ϖ_j.

The values of the parameters in PLV can be computed by using the Lagrangian multiplier method:

$$\Rightarrow \mu_{ij}=\frac{\left(\frac{1}{\log\left[\pi_i\times Pos(x_j;C_i)\right]}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c}\left(\frac{1}{\log\left[\pi_k\times Pos(x_j;C_k)\right]}\right)^{\frac{1}{m-1}}} \tag{4}$$

$$\Rightarrow c_i=\frac{\sum_{j=1}^{n}\mu_{ij}^{m}\times x_j}{\sum_{j=1}^{n}\mu_{ij}^{m}} \tag{5}$$

The mixing parameter: $\pi_i=\frac{1}{n}\sum_{j=1}^{n}\mu_{ij}$ (6)

μ_{ij} is employed to compute the fuzzy degree of memberships of the points to the clusters. π_i is used to compute the mixing parameters and c_i are updating the cluster centroids. The PLV clustering algorithm applies these steps iteratively until $\max\left|\mu_{ij}(t+1)-\mu_{ij}(t)\right|\le\varepsilon$.

III. PROPOSED EVALUATION METHOD FOR SCALING PARAMETER, β

This study proposes a new method for evaluating appropriate values for β.This method assumes a minimal degree of cost energy for samples which are close to any of the cluster centroids, i.e. samples which are legitimate cluster members. Consider the case where centroid c_m overlaps (or, is close to) sample x_n. The possibilistic function can be re-written as

$$Pos(x_n;C_m)=\frac{1}{1+\sum_{\substack{k=1\\k\neq m}}^{c}\exp\left[-\beta\times d^2(x_n,c_k)\right]}. \tag{7}$$

To ensure a valid clustering result, the possibility function must be larger than or equal to ρ, the minimal degree of the cost, where $\frac{1}{c}<\rho<1$. Therefore (3)

$$\frac{1}{1+\sum_{k=1,k\neq i}^{c}\exp\left[-\beta\times d^2(c_i,c_k)\right]}\ge\frac{1}{1+(c-1)\times\exp\left(-\beta\times d_{\min}^2\right)}\ge\rho \tag{8}$$

where $d_{\min}^2$ denotes the minimum of the square of the distances between all of the centroid pairs. In other words, β must satisfy

$$\Rightarrow\beta\ge\frac{-\log\left(\frac{\rho^{-1}-1}{c-1}\right)}{d_{\min}^2}. \tag{9}$$

In specifying $\beta=-\log\left[\left(\rho^{-1}-1\right)/(c-1)\right]/d_{\min}^2$, it follows that as the value of ρ increases, the "confidence" grows that a sample close to the centroid belongs to the corresponding population. Empirically, it is found that $\rho=0.99$ yields satisfactory results for most datasets.

IV. EXPERIMENTS

A. Lung Cancer Dataset

This dataset was published by Hong [10] and includes a total of 32 instances, each of which has 56 attributes with integer values ranging from 0 to 3. The instances are partitioned into three pathological lung cancer classes. It has been shown in previous studies that the RDA, KNN and Opt. Disc. schemes achieve a classification accuracy of 62.5%, 53.1% and 59.4%, respectively, when processing this dataset. Table 1 presents the accuracy of the three cancer classes and summarizes the optimal clustering results obtained by the PLV algorithm. The average accuracy is found to be 75.0%, which represents a significant

improvement relative to that achieved by the three schemes mentionesd above. Table 2 demonstrates the clustering results obtained by the PLV algorithm when applying the Annealing evaluation methods to the lung cancer database.

B. Dermatology Database

The second example concerns the diagnosis of erythemato-squamous dermatology diseases, and uses the dataset first proposed by Demiroz [11]. This dataset contains 366 instances, divided non-equally between the following six disease groups: psoriasis (112), seboreic dermatitis (61), lichen planus (72), pityriasis rosea (49), cronic dermatitis (52), and pityriasis rubra pilaris (20). Table 3 presented the optimal clustering results obtained using the PLV algorithm. The clustering results obtained by the PLV algorithm when using the proposed coefficient evaluation methods are summarized in Table 4.

V. CONCLUSION

In a future study, the PLV algorithm discussed in this study will be integrated with a semi-supervised learning mechanism in order to assist PLV in determining the number of clusters, *c*, in the dataset and to initialize the positions of the cluster centroids automatically such that a more accurate and efficient clustering performance can be obtained.

REFERENCES

[1] Z.-X Yin, and J.-H Chiang, "Patterns Discovery on Complex Diagnosis and Biological Data using Fuzzy Latent Variables," *2007 IEEE 23rd International Conference on Data Engineering*, *ICDE 2007*, Turkey, pp. 576-585, 2007.

[2] G. Beni, and X. Liu, "A Least Biased Fuzzy Clustering Method," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 16, no. 9, pp. 954-960, 1994.

[3] R. J. Hathaway, and J. C. Bezdek, "Switching Regression Models and Fuzzy Clustering," *IEEE Trans. On Fuzzy Systems*, vol. 1, no. 3, pp.195-203, 1993.

[4] J.-H. Chiang, and Z.-X. Yin, "Unsupervised Minor Prototype Detection using an Adaptive Population Partitioning Algorithm," *Pattern Recognition*, vol. 40, no. 11, pp. 3132-3145, 2007.

[5] N. B. Karayiannis, and P.-I Pai, "Fuzzy Algorithms for Learning Vector Quantization," *IEEE Trans. On Neural Networks*, vol. 7, no. 5, pp.1196-1211, 1996.

[6] R. Krishnapuram, and J. M. Keller, " A possibilistic Approach to Clustering", *IEEE Trans. Fuzzy System,* vol. 1, no. 2, pp 98-110, 1993.

[7] K. Rose, E. Gurewitz, and G. C. Foz, "Constrained Clustering as an Optimization Method," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no.8, pp. 785-794, 1993.

[8] T. A. Runkler, and J. C. Bezdek, "Alternating Cluster Estimation: A New Tool for Clustering and Function Approximation," *IEEE Trans. On Fuzzy System*s, vol. 7, no. 4, pp.377-393, 1999.

[9] H. Suh, J.-H. Kim, and F.C.-H Rhee, "Convex-Set-Based Fuzzy Clustering," *IEEE Trans. On Fuzzy Systems*, vol. 7, no. 3, pp.271-285, 1999.

[10] Z. Q. Hong, and J.Y. Yang, "Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane," *Pattern Recognition*, vol. 24, no. 4, pp. 317-324, 1991.

[11] G., H. Demiroz, A. Govenir, and N. Ilter, "Learning Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals," *Artificial Intelligence in Medicine*, Vol. 13, pp. 147-165, 1998.

Table 1 Accuracy of each class in lung cancer dataset obtained by PLV

	Assigned class		
	I	II	III
Accuracy	88.89%	53.85%	90.00%

Table 2 value of scaling coefficient obtained for lung cancer dataset with the proposed method

		Annealing method
Lung Cancer	β	1.1018
	accuracy	75.00%

Table 3 Accuracy of each class in dermatology database obtained by PLV algorithm

	Assigned class					
	I	II	III	IV	V	VI
Accuracy	99.11 %	70.49 %	100 %	95.92 %	100 %	100 %

Table 4 Clustering results obtained for dermatology dataset using the proposed evaluation methods

		Annealing method
Dermatology	β	1.8416
	accuracy	93.99%