A Bayesian Possibilistic C-Means Clustering Approach for Cervical Cancer Screening

Fang-Qi Li and Shi-Lin Wang[1]

School of Information Security Engineering,

Shanghai Jiaotong University, Shanghai, China

Abstract: Recently, more and more researchers and medical staffs have paid special attention to the treatment of the cervical cancer due to its high lethality and morbidity. Early screening of this kind of disease is of vital importance. In this paper, we propose an automatic cervical cancer screening algorithm that analyzes the related risk factors to provide some preliminary diagnostic information for doctors. Since a number of risk factors are considered as privacies, some patients refused to provide the corresponding information. Such severe amount of missing attributes leads to great difficulty for many automatic screening algorithms. To solve this problem, a Bayesian possibilistic C-means (BPCM in short) clustering algorithm is proposed to discover the representative patterns from the complete data and to estimate the missing values of a specific sample using its closest representative pattern. After data completion, a two-stage fuzzy ensemble learning scheme is proposed to derive the final screening result. In the first stage, the bootstrap aggregation (bagging in short) procedure is adopted to sample the entire class-imbalanced dataset into a number of class-balanced subsets. In the second stage, a number of weak classifiers are trained on each subset and a fuzzy logic based approach is designed to analyze the classification results of the weak classifiers and obtain the final class label. Experiments have been conducted on a dataset containing 858 patients. From the experiment results, it can be observed that the proposed BPCM can effectively discover the underlying patterns and is reliable in estimating the missing attribute compared with the traditional approaches. Moreover, by applying the proposed fuzzy ensemble learning scheme, the final classification results on the completed data by BPCM are promising (an accuracy of

76% or a positive sensitivity of 79%) under the severe missing-attribute scenario (only 6% samples with complete data).

Keywords: cervical cancer screening; Bayesian possibilistic C-Means clustering; fuzzy logic; ensemble learning; granular computing

[1] Corresponding author. Tel.: +8621-3420-5025; Fax: +8621-3420-5025.

Address: No. 800, Dong Chuan Rd., School of Information Security Engineering, Shanghai Jiaotong University, 200240, Shanghai, China.

E-mail addresses: wsl@sjtu.edu.cn (Shi-Lin Wang).

1. Introduction

With the past decades witnessing the blooming development of data science as well as bioscience, increasing efforts are now being devoted to combine the techniques in these two fields, and the results have been fruitful and inspiring [1][2][3][4]. Apart from applying the latest learning algorithms to biomedical data sets [5], many works were motivated by the unique challenges that biomedical data inherited from the biological and clinical circumstance. In the literature concerning bioscience and medical science, the challenges that are most frequently studied include the high dimensionality [6] that requires effective feature selection [7], severe class imbalance [8][9], and privacy issue [10][11] with the consequent uncertainty[12]. The above difficulties in biomedical data call for modifications to the classical machine learning algorithms as well as the data mining tools for better performance.

Among various data mining tasks in biomedical contexts, computer-aided diagnosis [25][26][27][29] has aroused special attention. The reason for its popularity is that computer-aided diagnosis systems can help to save lives in countries where medical source is still scanty, and such scantiness is confronting many low-income countries. Existing computer-aided diagnosis (CAD) approaches usually take advantage of machine learning. CAD utilizes algorithms to process complex data, including images and unstructured data. For example, in cancer screening, a CAD system has to process X-ray and other scanning images [57], together with other factual data. In a simplified setting, available data about correct diagnosis is inputted into some learning algorithms, which will learn the pattern and make predictions to help the doctors. Some classic learners are k Nearest Neighbours, naive Bayes, neural network, etc. [29][30]; however, appropriate modifications are necessary for traditional machine learning tools to perform well enough in biomedical context.

In recent years, cervical cancer has attracted much attention by being the fourth most common cause of death from cancer in women [34][35][36]. While there are good screening programs available in developed countries to lower the overall mortality, seventy percent of occurrence and ninety percent of death take place in developing countries [31]. Hence an auxiliary screening scheme based on easily

accessible factors is in urgent need. Researchers have demonstrated that some risk factors such as sexual history, smoking history, various symptoms of potential complications, etc. provide much useful information in cervical cancer screening and diagnosis [32]. However, due to privacy concerns and other reasons, not all the above information can be collected, which leads to great difficulties for the existing computer-aided diagnosis approaches. To solve this problem, Fernandes K. et al. in [27] proposed to use mean substitution to fix the missing attributes. But this imputation approach is too blunt to yield a satisfactory performance. In addition, the work done is mainly a study of sharing knowledge within the regression models instead of improving the specificity or the accuracy of classification in screening.

Some pioneering works [13][14][15] have demonstrated that granular computing (GrC) can neatly handle the uncertainty and vagueness in data mining tasks. Granular computing concerns processing collections of entities that are formed by similarity or indistinguishablity. So far, GrC, or fuzzy methodologies have found wide application in machine learning, ranging from the fundamental theory aspects to frontier applications [17][18], especially in cybernetics, expert systems, and biomedical environments[19][20][21] [22][23][24]. Their procedure usually includes generalizing a classical machine learning model to its fuzzy counterpart and demonstrating the improvement in performance. For instance, [19][21][24] generalize a fuzzy version of Support Vector Machine to fit specific data, while [20][22] make use of fuzzy logic to help to make decisions. Inspired by the granular computing (GrC) philosophy, a new algorithm is proposed in this paper to provide accurate and robust cervical cancer screening results. To handle the severe data incompleteness, a Bayesian version of Possibilistic C-Means Clustering algorithm is proposed that can detect some valuable patterns robustly for improved imputation. After data completion, a bagging scheme and an ensemble module is designed for classification with class imbalanced data. The major contributions of the proposed algorithm are three-folds: i) The proposed BPCM clustering can discover the underlying data distribution and extract meaningful representative patterns from limited complete data; ii) In BPCM, the actual number of patterns (i.e. the number of clusters) for the data is determined automatically; iii) A

fuzzy ensemble learning scheme is proposed, which can deal with the class-imbalance problem and handle the uncertainties in data collection, missing attribute completion, etc. Hence, a reliable cervical cancer screening result can be obtained. In a data set consisting of information provided by 858 patients, our framework can provide screening prediction with the accuracy of 76% and the sensitivity of 79%. Which are superior to the results obtained by established methods.

The rest of the paper is organized as follows. Section 2 specifies the challenges in the studied data set. Section 3 proposes the BPCM clustering algorithm. Section 4 describes the details of the bagging and fuzzy rule ensemble modules. Section 5 presents the experiment results, where attention is focused on the comparison between clustering schemes dealing with the missing attributes. Section 6 concludes the paper.

2. Challenges and Attempts

2.1. The missing attribute problem

Patient Number	Age	Number of pregnancies	Smoking status	Hormonal contraceptives usage	Number of STDs	Time since first STD diagnosis	Time since last STD diagnosis
1	44	N/A	True	False	0	N/A	N/A
2	41	4	False	True	1	21	21
3	36	3	N/A	True	0	N/A	N/A
4	34	3	False	N/A	N/A	N/A	N/A
5	36	2	False	True	0	N/A	N/A

Table 1 An example of missing attributes in the risk factor data set, where “N/A” denotes the missing attributes.

Although there have been CAD frameworks for cervical cancer based on visual information [58]. There remains one major problem in cervical cancer screening in the risk factor collection process. Since questionnaires concerning cervical cancer factors often involve queries on some private information such as “number of sexual

partners", "pregnancy status" and other gynecological diseases, very few participants are willing to provide all the related information. Moreover, the informative photographing methods might not always be feasible, thus a screening framework that not dependent on computer-vision data will have practical advantage in cost. For example, in the dataset [27][28], only 6% of participants provided complete data and most of the data lack at least two components. In the worst case, some participants provide almost no informative component. Hence, missing data overwhelm the data set of cervical cancer's risk factor provided by [27]. Among all the thirty-two features, the most frequent missing attributes are "the time since the first/last sexually transmitted disease (STD) diagnosis", which were ignored by over 90% of surveyed subjects. Around 12% of the subjects decided not to provide any information about their STD situation, which caused ten to thirteen missing attributes in their corresponding feature vectors, leaving little information to be exploited. Table 1 list some risk factors for several patients.

In order to deal with the missing attributes, three single-value imputation methods can be used [38]:

(1) *Mean imputation/substitution*: This kind of approaches use the average value of all the valid data of a specific attribute to fill the missing entries;

(2) *Regression imputation*: These approaches assume that data are subject to a linear/polynomial pattern. However, in cervical cancer screening, many attributes are of Boolean values, which hinders this kind of solutions;

(3) *Hot deck imputation*: By assuming a distance metric or a generative distribution over the data set, this family of approaches estimates the missing attributes by assigning a most probable value based on the inherent data distribution obtained from the complete data. Various kinds of clustering approaches have been widely used in hot deck imputation, including hard/crisp C-means (HCM) clustering [31], fuzzy C-means (FCM) clustering [33], etc. A brief review of the clustering approaches is provided in *Appendix I*. In the cervical cancer screening task, the hot deck imputation approaches are more popular because they can both estimate the missing value and provide informative knowledge about the inherent data distribution.

2.2. Missing attribute estimation based on data clustering

Fixing missing attributes by data clustering usually consists of the following two steps: i) Clustering on the complete data is performed and the converged cluster centroids are adopted as the representative patterns to depict the inherent structure of all the input data; ii) For any data with missing attributes, the closest centroid is found based on the known attributes, then missing value are filled with the corresponding components from that centroid. Generally speaking, the missing value estimation accuracy depends on the performance of the clustering approach.

For the cervical cancer dataset [27], finding the representative patterns from the complete data is a non-trivial task. The major difficulties can be summarized as follows: i) *Noise and uncertainty in data collection*: Owing to some subjective (e.g. misremembering certain information) and objective facts (e.g. slipping of the pen in the questionnaire), the collected data may have noise and uncertainty which brings obstructions for deterministic clustering approaches such as the Hard C-Means (HCM). Moreover, some noise and uncertainty in data will create outlier samples, which will prevent the clustering algorithms to discover the underlying representative patterns; ii) *Limited data*: The number of complete data counts for a mere 6% of the entire dataset, this limitation calls for highly robust clustering algorithms. Since most clustering methods, e.g. Fuzzy C-Means clustering (FCM) and Possibilistic C-Means clustering (PCM) adopt iterative search in their optimization procedure, how to derive robust and stable clustering results with various initializations becomes an important problem, especially when the number of samples for clustering is limited; iii) *Insufficient information about the underlying data distribution*: For most data clustering algorithms, the number of clusters should be preset before clustering and an inappropriate setting will greatly harm the performance [31]. However, for the risk factors related to the cervical cancer, information to infer the appropriate number of clusters is insufficient.

To overcome the difficulties mentioned above, a new data clustering algorithm based on the Bayesian theory and the fuzzy theory is proposed, which is named as

BPCM. The proposed BPCM has the following characteristics that help to handle the mentioned problems in estimating missing attributes. First, a soft clustering algorithm is designed to model the noise and uncertainty in the collected data. Specifically, to be robust against outliers, the possibilistic membership constraints in PCM is adopted to reduce the influence from outliers. Second, a Bayesian formulation is adopted to obtain the cluster centroids, i.e. the representative patterns, during which the number of clusters is automatically determined. This formulation makes BPCM robust against random initializations even with limited data. With the above two properties, the proposed BPCM is able to provide reliable representative patterns for the risk factors related to cervical cancer. In the next Section, a detailed description of the proposed BPCM algorithm will be presented.

3. The Proposed Bayesian Possibilistic C-Means (BPCM) Clustering

To extract representative patterns from the limited complete data, a Bayesian possiblistic C-Means clustering approach has been designed, which combines the merits of possibilistic membership constraints and Bayesian estimation. Notations of the frequently used variables are listed in Table 2.

Notation	Meaning
$\boldsymbol{x_i}$	A vector representing the *i*-th original observation.
$\boldsymbol{X}$	$\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x}_{N_s}\}$, where N_s is the number of samples.
χ	The space where the observation data lies, and $\mathbf{x}_i \in \chi$.
$\boldsymbol{c_j}$	The *j*-th centroids of $\mathbf{X}$.
$\boldsymbol{C}$	$\boldsymbol{C} = \{\boldsymbol{c_1}, \boldsymbol{c_2}, ..., \boldsymbol{c}_{N_c}\}$, where N_c is the number of clusters.
$u_{i,j}$	The membership value of the *i*-th observation belonging to the *j*-th cluster.
$\boldsymbol{u_i}$	$\boldsymbol{u_i} = (u_{i,1}, u_{i,2}, ..., u_{i,N_c})^{\mathrm{T}}$.
$\boldsymbol{U}$	$\boldsymbol{U} = \{\boldsymbol{u_1}, \boldsymbol{u_2}, ..., \boldsymbol{u}_{N_s}\}$.

d	$d:\chi\times\chi\to[0,\infty)$ is a measure of distance.

Table 2 Summary of mathematical notations

3.1. Derivation of the optimal cluster centroids

To obtain a robust estimation of the underlying representative patterns within the limited observation data, the Bayesian estimation is adopted. The Maximum Likelihood (ML) criterion is employed, which aims to maximize the conditional probability p($\boldsymbol{X}|\boldsymbol{C}$). By introducing the membership distribution $\boldsymbol{U}$, the conditional probability can be reformulated as,

$$p(\boldsymbol{X}|\boldsymbol{C}) = \int p(\boldsymbol{X},\boldsymbol{U}|\boldsymbol{C})d\boldsymbol{U} = \int p(\boldsymbol{U}|\boldsymbol{C})p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{U})d\boldsymbol{U}$$
$$= \int p(\boldsymbol{U}|\boldsymbol{C})\prod_{i=1}^{N_s} p(\boldsymbol{x_i}|\boldsymbol{C},\boldsymbol{u_i})\,d\boldsymbol{U} \quad (1)$$

where $p(\boldsymbol{U}|\boldsymbol{C})$ is a prior distribution and can be taken as a constant with uninformative prior applied [32]. Such kind of target function is usually optimized using the Expectation-Maximization (EM) routine [32]. In the E-step, the expectation of the latent variable $\boldsymbol{U}$ is computed, and in the M-step, the local-maxima of the log-likelihood is obtained with the membership distribution fixed. Similar to the definitions in the Gaussian mixture model (GMM) [32], we have:

$$p(\boldsymbol{x_i}|\boldsymbol{C},\boldsymbol{u_i}) = \frac{1}{Z(\boldsymbol{C},\boldsymbol{u_i})}\prod_{j=1}^{N_C}\exp\{-d(\boldsymbol{x_i},\boldsymbol{c_j})\}^{u_{i,j}} \quad (2)$$

In which

$$Z(\boldsymbol{C},\boldsymbol{u_i}) = \int\prod_{j=1}^{N_C}\exp\{-d(\boldsymbol{x_i},\boldsymbol{c_j})\}^{u_{i,j}}\,\mathrm{d}\boldsymbol{x_i} \quad (3)$$

Note that different from the settings in GMM, the constraint on $u_{i,j}$ is relaxed from one-hot coding to a general $\forall i,j{:}u_{i,j}\in[0,1]$. Computing the normalization term Z w.r.t $\boldsymbol{C}$ and $\boldsymbol{u_i}$ requires an integral over all possible value of $\boldsymbol{x_i}$ which is usually done by sampling methods.

In the E-step, the membership distribution over $\boldsymbol{U}$ can be regarded as the posterior distribution in and is computed as:

$$p(\boldsymbol{u_i}|\boldsymbol{x_i},\boldsymbol{C}) = \frac{\frac{1}{Z(\boldsymbol{C},\boldsymbol{u_i})}\prod_{j=1}^{N_C}\exp\{-d(\boldsymbol{x_i},\boldsymbol{c_j})\}^{u_{i,j}}\cdot p(\boldsymbol{u_i})}{\int\frac{1}{Z(\boldsymbol{C},\boldsymbol{v_i})}\prod_{j=1}^{N_C}\exp\{-d(\boldsymbol{x_i},\boldsymbol{c_j})\}^{v_{i,j}}\cdot p(\boldsymbol{v_i})\mathrm{d}\,\boldsymbol{v_i}} \quad (4)$$

where the integral in the denominator is also computed by sampling. Given the membership distribution in Eqn.4, the negative log likelihood in the M-step can be formulated as:

$$-\ln p\,(\boldsymbol{X}|\boldsymbol{U},\boldsymbol{C}) = \sum_{i=1}^{N_S}\sum_{j=1}^{N_C} u_{i,j}\cdot d\big(\boldsymbol{x}_i,\boldsymbol{c}_j\big) + f(\boldsymbol{C})$$

$$\approx \sum_{i=1}^{N_S}\sum_{j=1}^{N_C} u_{i,j}\cdot d\big(\boldsymbol{x}_i,\boldsymbol{c}_j\big) \tag{5}$$

Compared with the dominating term, the term $f(\boldsymbol{C})$ in Eqn.5 is neglectable (a brief proof is given in *Appendix II*). Then, the conditional log-likelihood $p(\boldsymbol{X}|\boldsymbol{U},\boldsymbol{C})$ is solely determined by $d(\boldsymbol{x}_i,\boldsymbol{c}_j)$ and gradient-based methods such as steepest gradient descent can be adopted to optimize Eqn.5 w.r.t $\boldsymbol{C}$.

In summary, the algorithm of the proposed BPCM is given in Algorithm. 1.

Algorithm 1 Bayesian Possibilistic C-Means Clustering

Input: $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N_S} \in \mathcal{X}^N$, N_C, $p(\boldsymbol{u})$, ϵ
Output: $\boldsymbol{C}$

Sample a sequence of $\{\boldsymbol{u}_k\}_{k=1}^{K}$ according to $p(\boldsymbol{u})$ in $[0,1]^{N_C}$;
Sample a sequence of $\{\boldsymbol{x}_l\}_{l=1}^{L}$ in $\mathcal{X}$;
Random initialize $\boldsymbol{C}^{(0)}$, $t = 0$;
while NOT ($t \geq 1$ AND $||\boldsymbol{C}^{(t)} - \boldsymbol{C}^{(t-1)}|| \leq \epsilon$) **do**
  for all $\boldsymbol{x}_i$ in $\boldsymbol{X}$ **do**
    Compute the posterior $p(\boldsymbol{u}_i|\boldsymbol{x}_i, \boldsymbol{C}^{(t)})$ in (4) by integrating on $\{\boldsymbol{u}_k\}_{k=1}^{K}$, during which $Z(\boldsymbol{C}^{(t)}, \boldsymbol{u}_i)$ has to be computed using (3), which can be obtained by integrating on $\{\boldsymbol{x}_l\}_{l=1}^{L} \cup \boldsymbol{X}$
    Substitute $u_{i,j}$ in(5) with their posterior expectations;
  end for
  Optimize w.r.t. $\boldsymbol{C}$ using gradient-based methods,save as $\boldsymbol{C}^{(t+1)}$;
  t++;
end while

Algorithm. 1 The pseudo-program of BPCM

To integrate out $\boldsymbol{x_i}$, a distribution over χ is necessary, which is approximated by $\boldsymbol{X}$ together with L extra samples. In practice, it is convenient to set $L = 0$, so we do not intentionally introduce additional samples in χ besides the provided data set. When sampling on the membership vector space, the sampling points are evenly distributed along each dimension.

3.2. Properties of BPCM

In this subsection, some properties of BPCM are presented to illustrate its

effectiveness in missing attribute estimation.

Property I: BPCM will converge properly after a number of iterations.

Proof: Formulating with a variational approach, the log-likelihood of the data set $\boldsymbol{X}$ w.r.t $\boldsymbol{C}$ is given by,

$$\ln p(\boldsymbol{X}|\boldsymbol{C}) = \int q(\boldsymbol{U}) \ln \left\{ \frac{p(\boldsymbol{X},\boldsymbol{U}|\boldsymbol{C})}{q(\boldsymbol{U})} \right\} \mathrm{d}\boldsymbol{U} + \mathrm{KL}(q(\boldsymbol{U})||p(\boldsymbol{U}|\boldsymbol{X},\boldsymbol{C})) \quad (6)$$

where q denotes an arbitrary distribution over the membership distribution $\boldsymbol{U}$ and $\mathrm{KL}(\cdot)$ denotes the ordinary Kullback-Leibler divergence between q and the posterior distribution over $\boldsymbol{U}$. Expanding the first term in Eqn.6, we have,

$$\int q(\boldsymbol{U}) \ln \left\{ \frac{p(\boldsymbol{X},\boldsymbol{U}|\boldsymbol{C})}{q(\boldsymbol{U})} \right\} \mathrm{d}\boldsymbol{U} = \mathrm{H}(q) + \mathrm{E}_q[\ln p(\boldsymbol{X},\boldsymbol{U},\boldsymbol{C})] \quad (7)$$

where $\mathrm{H}(\cdot)$ denotes the differential entropy and $\mathrm{E}_q[\cdot]$ is the expectation taken w.r.t the distribution q. In the E-step, the KL divergence is minimized, and in the M-step, the expectation is maximized, which guarantees the lower bound of the likelihood w.r.t $\boldsymbol{C}$ is monotonously increasing. Hence, BPCM will converge to a local optimum of $\ln p(\boldsymbol{X}|\boldsymbol{C})$ w.r.t $\boldsymbol{C}$.

Property II: BPCM is a generalized version of HCM, FCM, PCM, and GMM.

In Table 3, the major differences among HCM, FCM, PCM, GMM and BPCM are illustrated. According to the derivations in the previous subsection, sampling in the membership space is controlled by a prior distribution on $[0,1]^{N_c}$. Consider the membership space, the PCM and BPCM provide a maximum search space which includes the ones used in HCM, FCM, and GMM. Consider the E-step, the expectation over possible membership vectors in GMM and BPCM is a generalized version of the local maximization w.r.t the membership distribution in HCM, FCM, and PCM. Replacing the computation of the expectation by setting the membership to the value with the highest confidence yields the traditional local optimum solution.

Model	Membership space	E-step
HCM	One-hot($\{0,1\}^C$)	Local optimum
FCM	Regularized	Local optimum

	$(\{0,1\}^C \subset FCM's \subset [0,1]^C)$	
PCM	Irregularized ($[0,1]^C$)	Local optimum
GMM	One-hot($\{0,1\}^C$)	Expectation
BPCM	Irregularized ($[0,1]^C$)	Expectation

Table 3 Major differences among the five models.

Property III: BPCM is capable of avoiding the null membership assignment**.**

This property is equivalent to the statement that in the E-step of each iteration, a non-null membership assignment is more likely to be assigned a higher confidence. To formally address this statement, the following lemma is introduced:

Lemma: For every $\varepsilon > 0$ and $\boldsymbol{u}^* = \boldsymbol{u} + \varepsilon \cdot \mathbf{1}_j$, there exists δ, such that for any $\boldsymbol{x}$, $d(\boldsymbol{x},\boldsymbol{c_j}) < \delta$ implies $p(\boldsymbol{x}|\boldsymbol{u}^*) > p(\boldsymbol{x}|\boldsymbol{u})$, where $\mathbf{1}_j$ denotes a one-hot vector with its j-th component to be 1.

Proof: By applying the membership vectors $\boldsymbol{u}$ and $\boldsymbol{u}^*$ to Eqn.2 and Eqn.3, we have,

$$\frac{p(\boldsymbol{x}|\boldsymbol{u}^*)}{p(\boldsymbol{x}|\boldsymbol{u})} = \frac{Z(\boldsymbol{u})}{Z(\boldsymbol{u}^*)} \cdot \exp\{-d(\boldsymbol{x},\boldsymbol{c}_j)\}^{\varepsilon} \tag{8}$$

$$\frac{Z(\boldsymbol{u})}{Z(\boldsymbol{u}^*)} = \rho > 1 \tag{9}$$

where Eqn.9 is derived by comparing each term involved in the integral. Combining Eqn.8 and Eqn.9, it is suggested that by selecting any $\delta < \frac{\ln \rho}{\varepsilon}$, then for any $\boldsymbol{x}$, $d(\boldsymbol{x},\boldsymbol{c_j}) < \delta$, the inequality $p(\boldsymbol{x}|\boldsymbol{u}^*) > p(\boldsymbol{x}|\boldsymbol{u})$ holds. This finished the proof of the lemma. ■

Now we turn to the proof of Property III. Without loss of generality, we prove for the first component.

Proof: For the null membership assignment $\boldsymbol{u} = \mathbf{0}$ and a non-null assignment $\boldsymbol{u}' = (\varepsilon,0,0,\ldots,0)$, when the data sample is close to the first centroid $\boldsymbol{c_1}$, i.e. $d(\boldsymbol{x},\boldsymbol{c_1}) < \delta(\varepsilon = 1, \mathrm{j} = 1)$, we have $p(\boldsymbol{x}|\boldsymbol{u}') > p(\boldsymbol{x}|\boldsymbol{u})$. Since an uninformative prior on the

membership space is assumed, we have $p(\boldsymbol{u}) = p(\boldsymbol{u}')$. Combining this result with the Bayesian formula, the property is derived, i.e.

$$p(\boldsymbol{u}|\boldsymbol{x}) = \frac{p(\boldsymbol{u})p(\boldsymbol{x}|\boldsymbol{u})}{p(\boldsymbol{x})} < \frac{p(\boldsymbol{u})p(\boldsymbol{x}|\boldsymbol{u}')}{p(\boldsymbol{x})} = p(\boldsymbol{u}'|\boldsymbol{x}) \tag{10}$$

■

Property IV: BPCM can assign proper membership values for overlapping samples and outlier samples.

Proof: Without loss of generality, the total number of clusters is set to 2. By applying Property III iteratively with $\varepsilon = 1$ for both clusters we observe that an overlapping sample with distances to both centroids lower than a threshold $\min\{\delta(1,j)\}_{j=1,2}$ will have its confidence in $\boldsymbol{u} = \mathbf{1}$ higher than $\boldsymbol{u} = \mathbf{0}$, while an outlier with distance to both centroids larger than $\max\{\delta(1,j)\}_{j=1,2}$ will has the confidence in $\boldsymbol{u} = \mathbf{0}$ higher than $\boldsymbol{u} = \mathbf{1}$. ■

Remark: This is a remarkable distinction between BPCM and other models listed in Table 3. An outlier is an observation point that is distant from other observations and is more likely to be generated by noise [33]. An overlapping sample is an observation which is not an outlier but has similar distances to multiple cluster centroids. In data clustering, the outlier samples are preferred to be assigned very small membership values for all the clusters and the overlapping samples are preferred to be assigned large membership values to the clusters close enough and small membership values to the rest. In FCM and GMM, since the distances between the outlier samples and all the cluster centroids are similar (a large value), they will likely to be assigned a membership vector of $(\frac{1}{N_c}, \frac{1}{N_c}, ..., \frac{1}{N_c})^{\mathrm{T}}$ in the E-step, which will influence the estimation of the cluster centroids. Moreover, outlier samples and overlapping ones are indistinguishable.

To illustrate the membership assignment by BPCM for these two kinds of data points, we assume the $N_c = 2$ for simplicity and the synthetic butterfly dataset shown in Fig.1 is analyzed. The membership distributions of several data points by the proposed BPCM in the E-step are demonstrated with heatmap in Fig.2, where the

preference for possible weight is illustrated by hues.

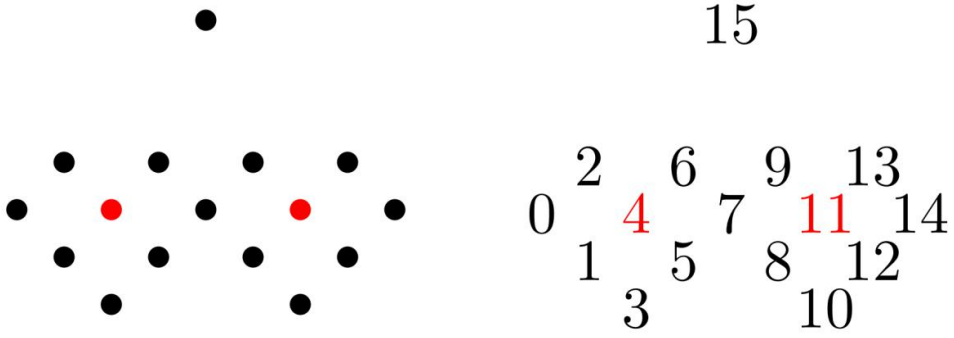


Fig.1 The indexed butterfly data set, where the 4-th and 11-th data points are the cluster centroids.

It is interesting to note that for the data point 7 and 15, their distance to the both centroids are the same, and thus FCM, PCM, and GMM would treat them equally in the E-step. However, the proposed BPCM distinguishes them by assigning the highest confidence to $\boldsymbol{u}_7 = (1,1)^{\mathrm{T}}$and $\boldsymbol{u}_{15} = (0,0)^{\mathrm{T}}$, since the grids (1,1) and (0,0) attracts the highest confidence in Fig.2(c) and (d) respectively. Hence, BPCM successfully distinguishes the overlapping samples and the outliers and assign appropriate membership values for them.

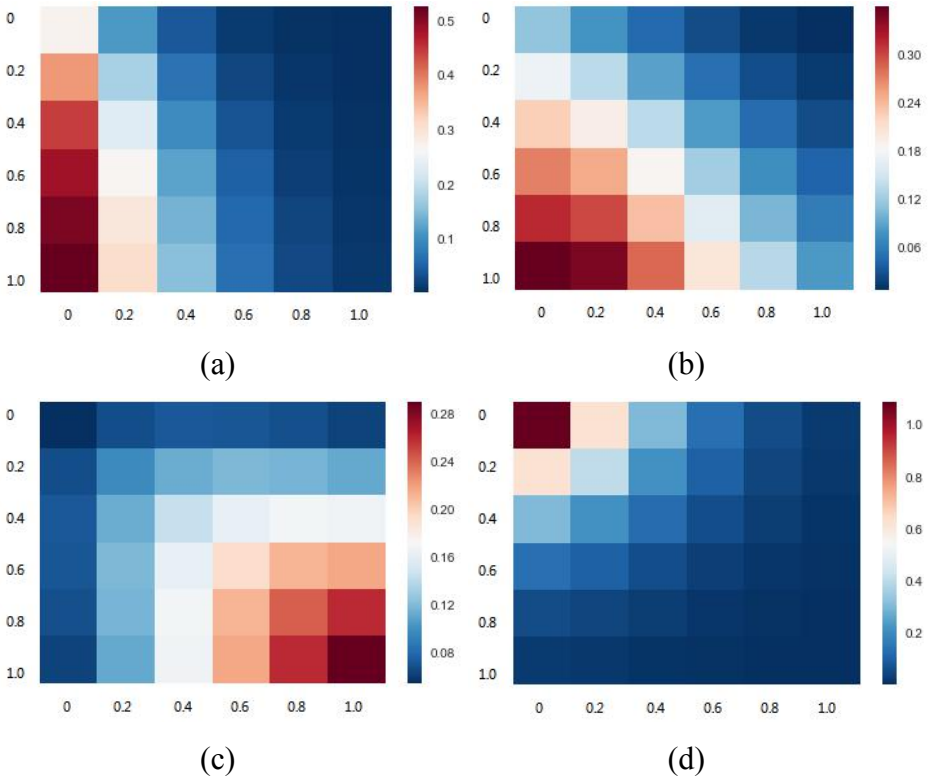


Fig.2 Heat maps that reflect the confidence assigned to different area in the

membership vector space. (a)-(d) denotes the assignment for points indexed 1,4,7,15, respectively. The vertical and horizontal axes denote the memberships to the clusters centered at the 4^{th} and the 11^{th} observation respectively.

3.3. Missing attribute estimation by BPCM

The missing attribute estimation procedure is illustrated in Fig.3 (Ω represents the data clustering algorithm adopted), which includes the following steps: i) Perform data clustering on the complete data subset and obtain the representative patterns (in the form of centroids); ii) For any data sample in the incomplete data subset, find its closest centroid based on the provided components in the sample vector. Assign the missing values with the corresponding component in the centroid; iii) Combining the complete data subset and the incomplete data subset after missing value estimation, the completed data set is derived. Note that the mean substitution can be covered by this framework by setting the number of clusters as one.

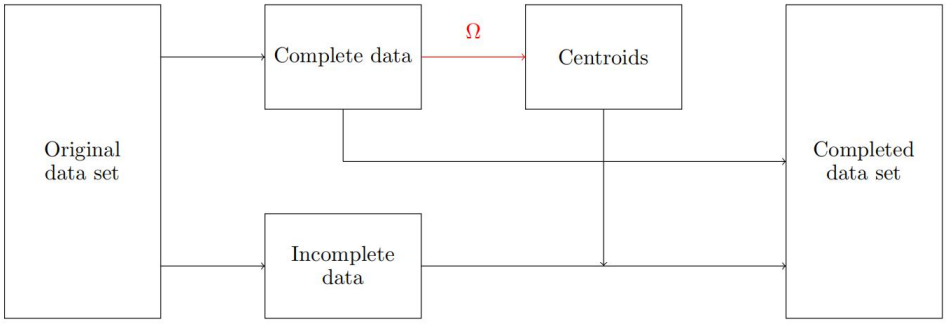


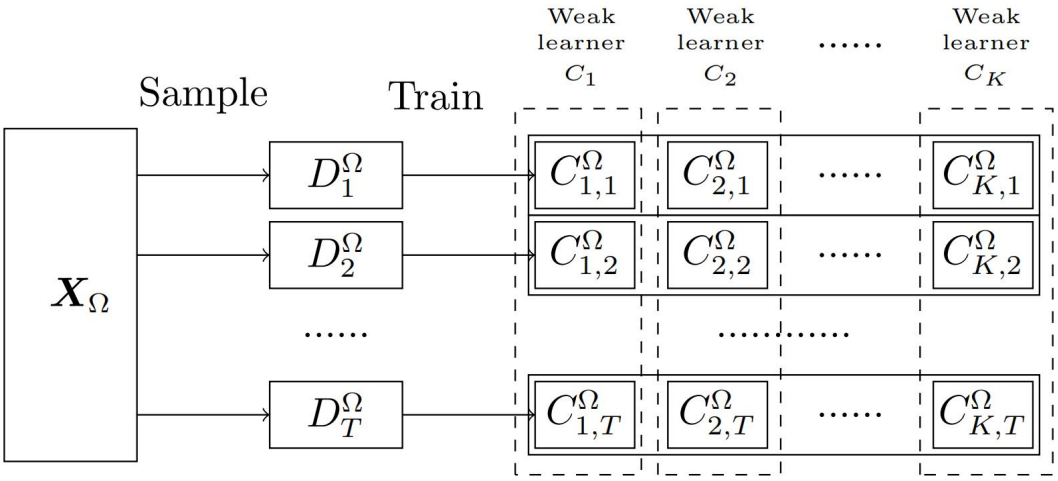
Fig.3 The flowchart of missing attributes estimation.

4. Classifier Design

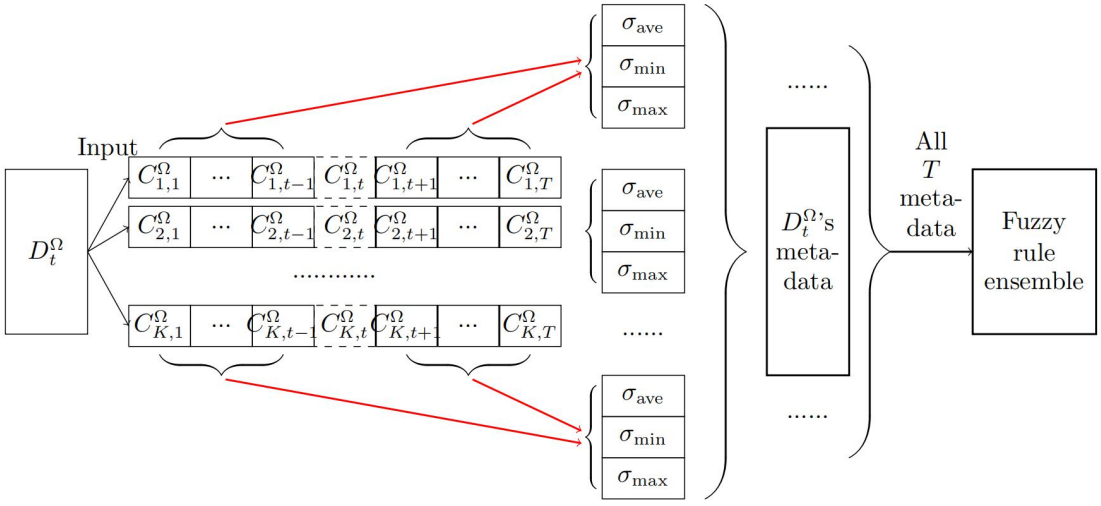
With the data completed, the cervical cancer screening reduces to a classification task, i.e., the patients are divided into two categories: positive (likely to have cervical cancer) or negative (unlikely to have cervical cancer). In this stage, there are still two major problems yet to be solved. First, in general, the data set is class imbalanced, i.e., for all the patients, the positive samples are much less than the negative samples. For example, in [27], the positive samples count for about 12% of the total samples. Directly training a classifier on such severely imbalanced data will not yield good

results. Second, the completed data contain much noise and uncertainty, which may be generated during the collection and missing data estimation stage. Hence, how to handle the uncertainty is another important issue in classifier design.

To solve the two problems above, a fuzzy ensemble learning approach is designed, which adopts a bagging strategy to generate class-balanced sub-datasets and a fuzzy logic based ensemble module to comprehensively analyze the classification results from each weak classifier and make the final decision. Denote the entire training set by $\boldsymbol{X}_{\Omega}$, where the subscript Ω represents the data clustering algorithm adopted. The proposed classification approach runs as follows and its flowchart is given in Fig.4.



(a) The bagging module



(b) The metadata generating module and the ensemble module

Fig.4. The bagging (a) and fuzzy rules ensemble learner (b), the red arrows indicate

that every input on the end is delivered to every receiver on the head.

i) Data sampling: As shown in Fig.4(a), a series of subsets with class-balanced samples are sampled from the entire training dataset $\boldsymbol{X}_\Omega$ and are denoted by $D_1^\Omega, D_2^\Omega, \ldots, D_T^\Omega$ (where T is the number of subsets and Ω denotes the algorithm used to fix the missing attributes). For 5-fold or 10-fold cross validation, in each subset D_t^Ω, 80% or 90% of the positive samples in $\boldsymbol{X}_\Omega$ are randomly selected to construct the positive constituents and similar number of negatives are selected from $\boldsymbol{X}_\Omega$ to construct the negative counterpart, which makes D_t^Ω a relatively balanced subset. Note that there is no overlapping negative sample in different subsets.

ii) Weak learner training: For each subset D_t^Ω $(1 \le t \le T)$, a series of weaker learners are trained on the data distribution in D_t^Ω, denoted by $C_{1,t}^\Omega, C_{2,t}^\Omega, \ldots, C_{K,t}^\Omega$ (where K is the number of weak learners).

iii) Meta-data generating: As shown in Fig. 4(b), for the k-th of weak learner, there are overall T classifiers trained from each subset, which are denoted by $C_{k,1}^\Omega, C_{k,2}^\Omega, \ldots, C_{k,T}^\Omega$. The bagging [50][51] algorithm is adopted to obtain the final estimation by combining the classification results from all the classifiers based on a specific combining strategy. Different combining strategy will result in different results. In the proposed approach, three kinds of combining strategies including maximum, minimum and average combining are adopted, denoted by $\sigma_{\max}(\cdot)$, $\sigma_{\min}(\cdot)$ and $\sigma_{\text{ave}}(\cdot)$, respectively. The meta-data for the samples in the t-th subset D_t^Ω based on the k-th weak learner is given by the combined classification outputs using all the combining strategies, i.e. $\sigma_{\max}, \sigma_{\min}, \sigma_{\text{ave}}\left(C_{k,1}^\Omega(x), \ldots, C_{k,t-1}^\Omega(x), C_{k,t+1}^\Omega(x), \ldots, C_{k,T}^\Omega(x)\right)$. Note that according to [49], when generating the metadata for a specific subset, the weak learner trained on itself should be excluded from the combining classifier set. Finally, for any observation in $(D_1^\Omega, D_2^\Omega, \ldots, D_T^\Omega)$, its meta-data contains $3 \times K$ combined classification results.

iv) Ensemble learning module by fuzzy logic: Inspired by [52][53], the fuzzy

IF-THEN rules filter is selected as the ensemble learner to handle the uncertainty in $\boldsymbol{X}_{\Omega}$. For any sample $\boldsymbol{x}$ in $\boldsymbol{X}_{\Omega}$, the fuzzy IF-THEN rule takes the form of "**Rule R: IF (** z_1 is A_1 **) and (** z_2 is A_2 **) and ... and (** z_V is A_V **) THEN** $\boldsymbol{x}$ **belongs to class** L_R **with a confidence** ***CF(R)***", where z_1 to z_v denote the specific components selected from $\boldsymbol{x}$'s meta-data, V is the number of components in $\boldsymbol{x}$'s meta-data considered ($V \leq 3K$). According to [37], in many cases, $V = 1$ yields a satisfactory result with high computational efficiency and thus this parameter setting is adopted in the proposed algorithm. A_1 to A_V are the corresponding antecedent fuzzy set for z_1 to z_v. Specifically, we use the five-partition form of antecedent fuzzy set. A strong partition with triangular fuzzy set is adopted, with the algebraic form of the membership function given in Eqn. 11 and its function curve in Fig.5.

$$A_1(z) = 1 - 4z, (0 \leq z \leq \frac{1}{4}), \quad A_2(z) = \begin{cases} 4z, (0 \leq z \leq \frac{1}{4}) \\ 2 - 4z, (\frac{1}{4} \leq z \leq \frac{1}{2}) \end{cases}, A_3(z) = \begin{cases} -1 + 4z, (\frac{1}{4} \leq z \leq \frac{1}{2}) \\ 3 - 4z, (\frac{1}{2} \leq z \leq \frac{3}{4}) \end{cases},$$

$$A_4(z) = \begin{cases} -2 + 4z, (\frac{1}{2} \leq z \leq \frac{3}{4}) \\ 4 - 4z, (\frac{3}{4} \leq z \leq 1) \end{cases}, \quad A_5(z) = -3 + 4z, (\frac{3}{4} \leq z \leq 1) \tag{11}$$

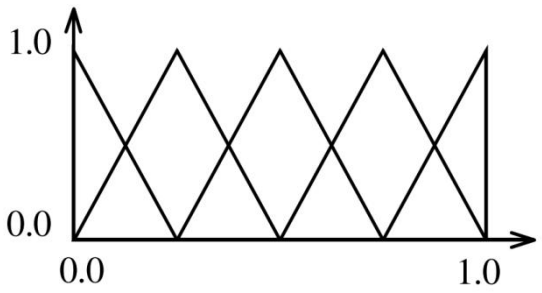


Fig.5 The antecedent fuzzy set membership function used.

The intensity of the condition of sample $\boldsymbol{x}$ following the rule R, denoted by $\mu_R(\boldsymbol{x})$, is computed by:

$$\mu_R(\boldsymbol{x}) = \prod_{v=1}^{V} A_v(\pi_v(\boldsymbol{x})) \tag{12}$$

where π is a projection operator that selects a specific component from the meta-data. Then the corresponding class label for rule R is determined by the overall

intensity of samples in both classes, i.e.,

$$L_R = \underset{l\in\{0,1\}}{\operatorname{argmax}} \{ \sum_{\boldsymbol{x} \in \text{Class } l} \mu_R(\boldsymbol{x}) \} \tag{13}$$

Intuitively, Eqn.13 measures the level that a rule is supported by a class l (l=0/1 denotes the positive/negative class, respectively) and the rule R is assigned with the class label L_R with a higher overall support.

The confidence for rule R is measured by how firmly R yields its answer, where the entropy formulation is adopted:

$$CF(R) = 1 + \sum_{l=0,1} \frac{\sum_{\boldsymbol{x} \in \text{Class } l} \mu_R(\boldsymbol{x})}{\sum_{\boldsymbol{x}} \mu_R(\boldsymbol{x})} \cdot \log\left(\frac{\sum_{\boldsymbol{x} \in \text{Class } l} \mu_R(\boldsymbol{x})}{\sum_{\boldsymbol{x}} \mu_R(\boldsymbol{x})} \right) \tag{14}$$

Then the discriminative power of the rule R is measured by a score formulated as:

$$\text{Score}(R) = CF(R) \cdot \sum_{\boldsymbol{x} \in \text{ Class } L_R} \mu_R(\boldsymbol{x}) \tag{15}$$

Finally, top H rules with the highest scores are kept into the final ensemble classifier. Note that for any class l, at least one rule with $L_R = l$ should be included to guarantee that all the labels can be assigned, and thus $H \geq 2$ in our case.

For a test sample $\boldsymbol{x}_{test}$ with completed attributes, the output of every weak learner (there are overall $K \times T$ weak learners) is computed. Using the three kinds of combining strategy, the meta-data for the test-sample can be derived, which contains 3 $\times K$ components. For each rule R in the H rules with the highest discriminative power, the sample's consistency to the rule is calculated as $\mu_R(\boldsymbol{x_{test}}) \cdot CF(R)$. Then the test sample is assigned a class label L_{Ropt}, where R_{opt} is the rule with the highest consistency.

5. Experiments and Discussions

5.1. Experiment setups

To evaluate the performance of the proposed algorithm in cervical cancer screening, the dataset in [27] is adopted as the studied target. For each patient in the dataset, there are overall twenty-seven risk factors recorded, including age, sexual experience (3 factors), smoking history (2 factors), intrauterine device (IUD) and hormonal contraceptives usage (2 factor), sexually transmitted disease (STD) related (13 factors), STD diagnosis time (2 factors), and previous cervical diagnosis (4

factors). Among all the risk factors, sixteen factors are Boolean variables and the rest are real variables. Before the subsequent processing, all the real variables are normalized by a linear transform to ensure that 95% of the values fall in the range [0,1].

In the classification stage, two kinds of metrics are adopted to evaluate the classification performance: the accuracy and the positive sensitivity, which are defined as follows,

$$\text{accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}} \tag{16}$$

$$\text{positive-sensitivity}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{17}$$

Where TP/TN/FN/FP are entries of the confusion matrix: true-positive, true-negative, false-negative and false-positive. The sick patients are treated as positive samples. So TP denotes the number of patients who actually has cervical cancer (according to four gold-standard benchmarks) and is classified as an infested one, analogously for other three entries. It should be noted that in our experiment, 80% of samples in the dataset/sub-dataset are used for training the classifier and the remaining 20% of samples are used for testing. The same experiment is repeated for five times with randomly selection of the training set and the average results are recorded.

5.2. Evaluations on Missing Attribute Estimation

To comprehensively evaluate the performance of the proposed BPCM algorithm in missing attribute estimation, the mean substitution and the FCM algorithm were used for comparison. Note that the estimation results by the PCM and HCM algorithms were unstable, i.e., they varied much with different initializations. Hence, they were excluded in the subsequent classification procedure. In addition, the total number of clusters in BPCM was set to eight and the converged number of clusters was four. It is interesting to note that in FCM, setting the number of clusters to four resulted in the best performance among other possible choices of it. This demonstrates that the proposed BPCM algorithm can automatically determine the correct number of

patterns of the data analyzed.

To evaluate the missing attribute estimation performance, all the samples in the complete dataset is divided into two sets, i.e. the training set and the test set. The training set contains 80% samples and it is used to provide the required information for the mean substitution, FCM and BPCM algorithms. The test set contains the remaining 20% samples and to simulate the samples with missing values. A number of attributes (from 10% to 50% with a step of 10%) for each sample in the test set are randomly erased. Then the missing attribute estimation performance is measured by the mean squared error (MSE) between the estimated values and their corresponding ground truths. As in Fig.6, the procedure is repeated for five times as a 5-fold cross validation, the average result measured in MSE and the running time is reported.

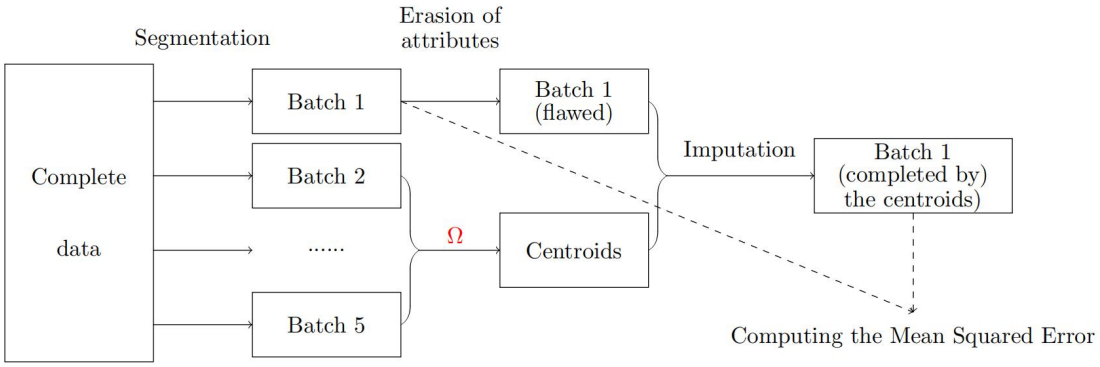


Fig.6. The flowchart of the measuring of imputation effect.

Fig. 7 shows the missing attribute estimation performance with the three approaches investigated. From the figure, it is observed that: i) The average MSE increases monotonically with the percentage of missing values for all the three approaches investigated; ii) The mean substitution approach performs well for samples with fewer missing attributes and the performances drops drastically with the increase of missing attributes. It is mainly because without adequate information, the estimation based on a simple averaging operation cannot be accurate and robust. In such scenario, the clustering based algorithm can provide better estimation results; iii) The proposed BPCM algorithm can always achieve the lowest MSE, which has demonstrated its effectiveness in missing attribute estimation.; iv) Compared with other two methods, BPCM has the longest running time due to the exhaustive search

in membership space with time complexity $O(N_s N_C(\frac{1}{\Delta})^{N_C})$ exponential to the number of clusters in each iteration, where Δ is the sampling step along each dimension. However, this could be reduced by applying Monte-Carlo Markov Chain methods instead of uniform sampling.

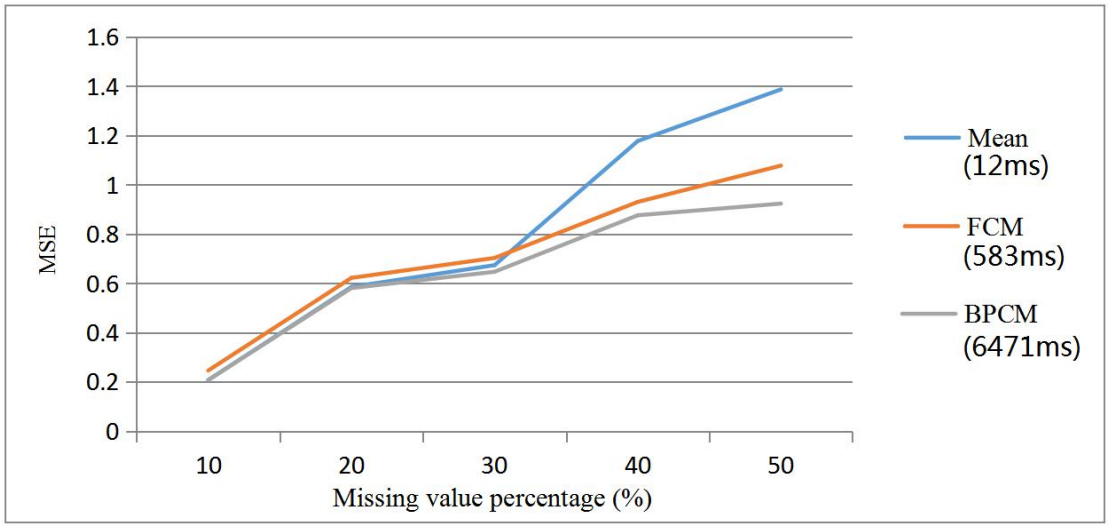


Fig. 7 The missing attribute estimation performance by the three approaches investigated.

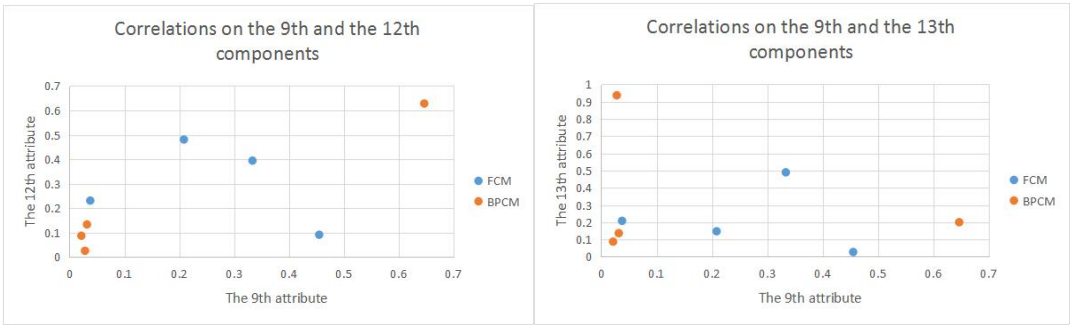
To show that the proposed BPCM can discover the underlying data structure and extract meaningful patterns, the following instance was analyzed. Considering the three risk factors including "STDs: condylomatosis" (the 9th component), "STDs: vulvo-perineal condylomatosis" (the 12th component), "STDs: syphilis" (the 13th component), the converged centroids of the FCM and BPCM are given in Table 4. To delve into the differences, we project the centroids onto each pair of attributes, the results are given in Fig.8.

From Table 4 and Fig.8, it is observed that the BPCM has learned the following useful patterns among the attributes. First, the ninth and the twelfth attribute have a strong correlation as shown in Fig.8 (a), which implies that in the dataset, if a patient was infected by condylomatosis, the specific category was very likely to be the vulvo-perineal condylomatosis; Second, as shown in Fig.8 (b) and (c), the ninth/twelfth attribute also has a strong relationship with the thirteenth attribute. For BPCM, the projected points fall close to $(0,0)$, $(0,1)$ and $(1,0)$, since these three attributes are innately boolean ones, the figure indicates a mutually exclusive relationship. That is

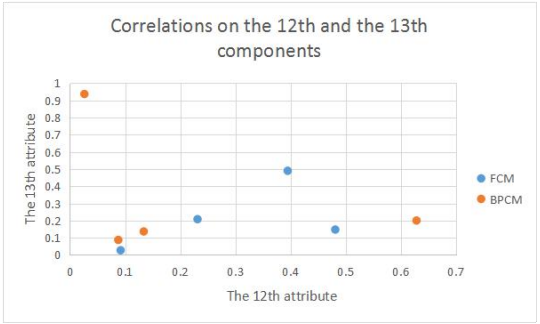
to say, if a patient was infected by condylomatosis, she was very unlikely to be infected by syphilis and vice versa. However, such patterns are not observed in the centroids extracted by FCM, as in Table 4, FCM tends to yield a mediorce result that contains less decisive information for boolean attributes.

Model	Centroid	9th component	12th component	13th component
FCM	1	0.21	0.48	0.15
	2	0.46	0.09	0.03
	3	0.33	0.39	0.49
	4	0.04	0.23	0.21
BPCM	1	0.65	0.63	0.20
	2	0.03	0.03	0.94
	3	0.02	0.09	0.09
	4	0.03	0.13	0.14

Table 4. Converged centroids by FCM and BPCM



(a) (b)



(c)

Fig.8 Illustration of the centroids extracted by FCM and BPCM.

To examine the correctness of the above patterns explored by BPCM, we have browsed the dataset and computed the following conditional probabilities:

$$p((\boldsymbol{x}(9),\boldsymbol{x}(12)) = (1,1)|(\boldsymbol{x}(9),\boldsymbol{x}(12)) \neq (0,0)) = 97.7\%$$
$$p(\boldsymbol{x}(9) = 0|\boldsymbol{x}(13) = 1) = 93.7\%, p(\boldsymbol{x}(12) = 0|\boldsymbol{x}(13) = 1) = 93.7\% \quad (18)$$
$$p(\boldsymbol{x}(13) = 0|\boldsymbol{x}(9) = 1) = 97.7\%, p(\boldsymbol{x}(13) = 0|\boldsymbol{x}(12) = 1) = 97.7\%$$

where $\boldsymbol{x}(m)$ denotes the m-th component of $\boldsymbol{x}$. Eqn. 18 indicates that the data distribution is in accordance with the patterns that BPCM discovered.

5.3. Classification Performance of the Weak Learners

The performance of the weak learners were measured in a 10-fold cross validation procedure. The dataset with the completed attributes was firstly separated into ten subsets. For each fold, nine subsets are merged and sampled into a balanced training set. Then it is input into weak learners. The result of classification of the test set was recorded. To avoid the variances caused by selection of the training samples, the entire procedure was repeated for five times and there were overall 5×10=50 sets of experiments considered.

During each test, three kinds of weak learners (K=3), including the naive Bayes classifier [54] with continuous variables, the k-nearest-neighbour [55] with k=5, and the logistic regression algorithms [56] were adopted. Each weak learner is trained on nine training sets and tested on the remaining set. Since classification at this stage was operated on a balanced data set, only the accuracy metric is adopted for evaluation.

(Average/Optimal)	Mean	FCM	BPCM
Naive Bayes	**63%/66%**	57%/65%	61%/65%
kNN (k=5)	59%/71%	63%/71%	**64%/80%**
Logistic regression	54%/59%	57%/**64%**	**60%**/62%

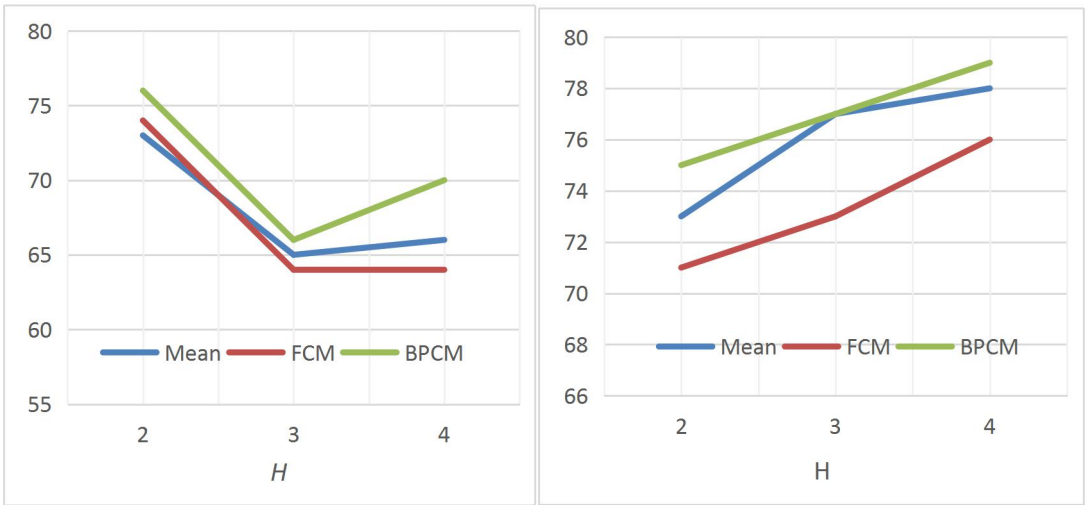
Table 5. Average and optimal performance of basic classifiers on data subsets.

Table 5 compares the classification performance using a single weak learner on the completed data with the three kind of missing attribute estimation approaches. Each entry in Table 5 is a summarized result from fifty experiments. From the table, it is observed that for the naïve Bayes classifier, the mean substitution approach

achieves the best performance. We confer that it is due to the reason that the naïve Bayes algorithm classifies a sample by comparing its distance to the centers of two classes. For the clustering based missing attribute estimation approaches, the completed data are gathered around multiple (four) centroids, which may confuse the naïve Bayes classification algorithm. Even so, the classification accuracy by BPCM are comparable to that obtained by mean substitution. On the other hand, for the kNN and logistic regression, the performance on the completed data by the BPCM algorithm usually outperforms those on the data set completed by the other two approaches. The highest classification result on the class-balanced subsets was obtained by the kNN algorithm on the BPCM completed dataset.

5.4. Classification Performance by the Fuzzy Ensemble Learning Algorithm

In this subsection, the classification performance using the fuzzy ensemble learning algorithm is evaluated. To select an appropriate number of rules, various choices of H was examined and the classification performance in terms of the accuracy and positive-sensitivity is given in Fig.9 (a) and (b) respectively.



(a) The accuracy(%) to H (b) The sensitivity(%) to H

Fig.9 Classification performance by the fuzzy ensemble leaner using various selections of H.

From the figure, the following points can be observed: i) there is a trade-off

between the accuracy and positive-sensitivity and thus for different kind of application, different number of rules should be adopted. In general high positive-sensitivity is preferred because we would like to discover the potential cervical cancer patient as early as possible and H=4 is the most proper setting; ii) the missing value estimation by BPCM always outperforms the other two algorithms investigated, which demonstrated the effectiveness of the proposed approach; iii) The proposed approach can achieve a relatively high accuracy and sensitivity in cervical cancer screening.

6. Conclusion

The cervical cancer is a high-death-rate disease threatening women's health. Computer-aid diagnosis systems for early detection of this disease without the help of experienced doctors is in urgent demand especially in developing countries. However, due to the privacy concerns and noise in data collection, the related risk factors obtained from the questionnaires usually contain much uncertainty, which brings great difficulty for accurate and robust diagnosis. To solve this problem, a complete solution based on the fuzzy theory is proposed in this paper and it can well handle the severe uncertainty in data collection. A new kind of fuzzy clustering algorithm, i.e. the BPCM, is proposed to extract the representative patterns from the limited complete data for missing attribute imputation. Then, a fuzzy ensemble learning scheme is designed to learn the inherent rules between the completed risk factor and the class label (positive or negative) under a high level of uncertainty. Experiment results on a dataset with 858 patients have shown the effectiveness of the proposed solution. The proposed approach can achieve an accuracy of 76% and a sensitivity of 79% in cervical cancer screening.

ACKNOWLEDGEMENT

The work described in this paper is fully supported by NSFC Fund (No. 61771310).

Reference

[1] Ohno Machado, Lucila. "Data science and informatics: when it comes to biomedical data, is there a real distinction?" *Journal of the American Medical Informatics Association*,20.6(2013):1009-1009.

[2] Adam, Nabil R., Robert Wieder, and Debopriya Ghosh. "Data science, learning, and applications to biomedical and health sciences." *Annals of the New York Academy of Sciences* 1387.1(2017):5-11.

[3] Dariusz Mrozek, Pawel Kasprowski, Bożena Małysiak-Mrozek, Stanisław Kozielski, "Life Sciences Data Analysis", *Information Sciences*, 384 (2017):86-89.

[4] Kerstin Denecke, Wolfgang Nejdl, "How valuable is medical social media data?" Content analysis of the medical web, *Information Sciences*,179.12 (2009):1870-1880.

[5] Baldi, and Pierre. "Deep Learning in Biomedical Data Science." *Annual Review of Biomedical Data Science* 1.1(2018):annurev-biodatasci-080917-013343.

[6] Tabik, S., Garzón, E. M., García, I., & Fernández, J. J. "High performance noise reduction for biomedical multi- dimensional data." *Digital Signal Processing* 17.4(2007):724-736.

[7] Oduntan, I. O., Toulouse, M., Baumgartner, R., Bowman, C., Somorjai, R., & Crainic, T. G. "A multilevel tabu search algorithm for the feature selection problem in biomedical data." *Computers and Mathematics with Applications* 55.5(2008):1019-1033.

[8] Min Su Lee, Sangyoon Oh, and Byoung-Tak Zhang. "Ensemble Learning Based on Active Example Selection for Solving Imbalanced Data Problem in Biomedical Data," *Proceedings of 2009 IEEE International Conference on Bioinformatics and Biomedicine*, Washington, DC, (2009): 350-355.

[9] Oh, Sangyoon, Min Su Lee, and Byoung-Tak Zhang. "Ensemble Learning with Active Example Selection for Imbalanced Biomedical Data Classification." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*

8.2(2011):316-325.

[10] Malin, Bradley A., Khaled El Emam, and Christine M. O'keefe.. "Biomedical data privacy: problems, perspectives, and recent advances." *Journal of the American Medical Informatics Association* 20.1(2013):2-6.

[11] Yang Yang, Xiang-han Zheng, Wen-Zhong Guo, Xi-Meng Liu, and Victor Chang. "Privacy-preserving Smart IoT-based Healthcare Big Data Storage and Self-adaptive Access Control System." *Information Sciences* (2018), in Press.

[12] Manuel Jiménez, Isaac Triguero, Robert John, "Handling Uncertainty in Citizen Science Data: Towards an Improved Amateur-based Large-scale Classification", *Information Sciences*, 479(2018):301-320.

[13] Witold Pedrycz, "Granular computing: an introduction," *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*, Vancouver, BC, Canada, 3(2001): 1349-1354.

[14] Yi-Yu. Yao. "On Modeling Data Mining with Granular Computing." *Proceedings of 25th Annual International Computer Software and Applications Conference. COMPSAC 2001*, Chicago, IL, USA, (2001):638-643.

[15] Witold Pedrycz. "Granular Computing: An Emerging Paradigm." Heidelberg Physica-Verlag (2001).

[16] Bezdek, and C. James. "Pattern Recognition with Fuzzy Objective Function Algorithms." *Advanced Applications in Pattern Recognition* 22.1171 (1981): 203-239.

[17] Adam Gacek, "Granular modelling of signals: A framework of Granular Computing", *Information Sciences*, 221(2013):1-11

[18] Andrzej Skowron, Jarosław Stepaniuk, Roman Swiniarski, "Modeling rough granular computing based on approximation spaces", *Information Sciences*, 184.1(2012):20-43

[19] Yu-Chun Tang, and Yan-Qing Zhang. "Granular support vector machines with data cleaning for fast and accurate biomedical binary classification." *Proceedings of 2005 IEEE International Conference on Granular Computing*, Beijing, 1(2005): 262-265.

[20] Xiao-Hua Hu, Guang-Rong Li, Illhoi Yoo, Xiao-Dan Zhang, Xu-Heng Xu. "A Semantic-based Approach for Mining Undiscovered Public Knowledge from Biomedical Literature." *Proceedings of 2005 IEEE International Conference on Granular Computing*, Beijing, 1(2005): 22-27.

[21] Xiu-Juan. Chen, R. Harrison, and Yan-Qing. Zhang. "Fuzzy support vector machines for biomedical data analysis." *Proceedings of 2005 IEEE International Conference on Granular Computing*, Beijing, 1(2005):131-134.

[22] Virant-Klun, and J. Virant. "Fuzzy Logic Alternative for Analysis in the Biomedical Sciences." *Comput Biomed Res* 32.4(1999):305-321.

[23] Harun Uğuz, Ali Öztürk, Rıdvan Saraçoğlu, Ahmet Arslan "A biomedical system based on fuzzy discrete hidden Markov model for the diagnosis of the brain diseases." *Expert Systems with Applications* 35.3(2008):1104-1114.

[24] Bo Jin, Y.C. Tang, Yan-Qing Zhang. "Support vector machines with genetic fuzzy feature transformation for biomedical data classification", *Information Sciences*, 177.2(2007):476-489

[25] Viergever, Max A, B. H. Romeny, and J. B. Van Goudoever . "Computer-aided diagnosis in chest radiography: a survey. " *IEEE Transactions on Medical Imaging* 20.12(2001):1228-41.

[26] Doi, Kunio. "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential." *Computerized Medical Imaging and Graphics* 31.4-5(2007):198-211.

[27] Fernandes K., Cardoso J.S., Fernandes J. “Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.” *Proceedings of 2017 Iberian Conference on Pattern Recognition and Image Analysis, Lecture Notes in Computer Science*, 10255(2017):243-520.

[28] Fernandes K., Cardoso J.S., Fernandes J. “Temporal Segmentation of Digital Colposcopies.” *Proceedings of 2015Iberian Conference on Pattern Recognition and Image Analysis, Lecture Notes in Computer Science,* 9117(2015):262-271.

[29] Antal, Blint. and A. Hajdu. "An ensemble-based system for automatic screening of diabetic retinopathy." *Knowledge-Based Systems* 60.2(2014):20-27.

[30] Kononenko, I. "Machine learning for medical diagnosis: history, state of the art and perspective. " *Artificial Intelligence in Medicine* 23.1(2001):89-109.

[31] Chiang, Mark Ming-Tso, and Boris Mirkin. "Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads." *Journal of Classification* 27.1(2010):3-40.

[32] Bishop, C. M. "Pattern Recognition and Machine Learning (Information Science and Statistics)". Springer-Verlag New York, Inc. 2006.

[33] Grubbs, and E. Frank. "Procedures for Detecting Outlying Observations in Samples." *Technometrics* 11.1(1969):1-21.

[34] Nam Phuong TranChien-Fu Hung, Richard Roden T.-C. Wu. "Control of HPV infection and related cancer through vaccination". Viruses and Human Cancer. Springer Berlin Heidelberg, (2014).

[35] Levin, B. "World Cancer Report 2008." *World Health Organization* (2015).

[36] Bosch, FX; de Sanjosé, S. "The epidemiology of human papillomavirus infection and cervical cancer". *Disease Markers*. 23.4(2007): 213–27.

[37] Tien Thanh Nguyen, Mai Phuong Nguyen, Xuan Cuong Pham, Alan Wee-Chung Liew. "Heterogeneous classifier ensemble with fuzzy rule-based meta learner", *Information Sciences*, 422(2018):144-160.

[38] Dan Li., Deogun J., Spaulding W., Shuart B. "Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method." *Proceedings of 2004 International Conference on Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science*, 3066(2004):573--579

[39] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin.: "Maximum likelihood from incomplete data via the EM algorithm." *J. of Royal Statistical Society Series* 39 (1977) 1–38

[40] Timm, Heiko, Christian Döring, and R. Kruse. "Different approaches to fuzzy clustering of incomplete datasets." *International Journal of Approximate Reasoning* 35.3(2004):239-249.

[41] Dunn, and C. Joseph. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters." *Journal of Cybernetics*

3.3(1973):32-57.

[42] Krishnapuram, R, and J. M. Keller. "A possibilistic approach to clustering." *IEEE Transactions on Fuzzy Systems* 1.2(1993):98-110.

[43] Krishnapuram, R, and J.M. Keller. "The possibilistic C-means algorithm: insights and recommendations." *IEEE Transactions on Fuzzy Systems* 4.3(1996):385-393

[44] A.Schneider, "Weighted possibilistic c-means clustering algorithms," *Proceedings of the Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000 (Cat. No.00CH37063)*, San Antonio, TX, USA, 1(2000):176-180.

[45] Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek. "A Possibilistic Fuzzy c-Means Clustering Algorithm." *IEEE Transactions on Fuzzy Systems* 13.4(2005):517-530.

[46] Forero, Pedro A., V. Kekatos, and G. B. Giannakis. "Robust Clustering Using Outlier-Sparsity Regularization." *IEEE Transactions on Signal Processing* 60.8(2012):4163-4177.

[47] Jörg Sander, Martin Ester, Hans-Peter Kriegel, Xiao-Wei Xu. "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications." *Data Mining and Knowledge Discovery* 2.2(1998):169-194.

[48] Siamak Mehrkanoon, Carlos Alzate, Raghvendra Mall, Rocco Langone, and Johan A. K. Suykens. "Multiclass Semisupervised Learning Based Upon Kernel Spectral Clustering," in *IEEE Transactions on Neural Networks and Learning Systems*, v26.4(2015):720-733.

[49] Benhur, Asa. "Support Vector Clustering." *Journal of Machine Learning Research* 2.2(2001):125-137.

[50] Breiman, and Leo. "Bagging predictors." *Machine Learning* 24.2(1996):123-140.

[51] Chun-Xia Zhang, R.P.W. Duin. "An experimental study of one-and-two-level classifier fusion for different sample sizes", *Pattern Recognit. Lett.* 32 (2011) 1756–1767.

[52] M.P.Sesmero, A.I. Ledezma, A. Sanchis, "Generating ensembles of heterogeneous classifiers using stacked generalization", *Wiley Interdiscip. Rev*. 5 .1(2015) 21–34.

[53] Ishibuchi, Hisao, and T. Yamamoto "Comparison of Heuristic Criteria for Fuzzy Rule Selection in Classification Problems." *Fuzzy Optimization and Decision Making* 3.2(2004):119-139.

[54] Rish I. "An empirical study of the naive Bayes classifier". *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 3.22(2001): 41-46.

[55] Cover, Thomas M., and Peter E. Hart. "Nearest neighbor pattern classification." *IEEE transactions on information theory.* 13.1 (1967): 21-27.

[56] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. "Applied logistic regression." Vol. 398. John Wiley & Sons, 2013.

[57] Rangayyan, Rangaraj M., Fabio J. Ayres, and JE Leo Desautels. "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs." *Journal of the Franklin Institute* 344.3-4 (2007): 312-348.

[58] Wen-Jing Li, Viara Van Raad, Jia Gu, Ulf Hansson, Johan Hakansson, Holger Lange, Daron Ferris. "Computer-Aided Diagnosis for Cervical Cancer Screening and Diagnosis: A New System Design in Medical Image Processing" *Proceedings of International Workshop on Computer Vision for Biomedical Image Applications* (2005):240-250.

Appendix I: A Review of Classic Clustering Algorithms

The ordinary model-based clustering algorithms minimize a loss function of the form:

$$L(\boldsymbol{X},\boldsymbol{U},\boldsymbol{C}) = \sum_{i=1}^{N_S}\sum_{j=1}^{N_C} u_{i,j}\cdot d\left(\boldsymbol{x}_i,\boldsymbol{c}_j\right) \quad \text{(A1)}$$

The joint optimization of $\boldsymbol{U}$ and $\boldsymbol{C}$ makes a global optimum unattainable, thus a two-step iterative optimization is used, where $\boldsymbol{U}$ and $\boldsymbol{C}$ are optimized in turn: while when $\boldsymbol{U}$ is optimized, $\boldsymbol{C}$ is fixed and vice versa.

The essential difference between some mainstream model-based clustering algorithms is the constraint exerted upon $\boldsymbol{U}$, which can be summarized in Table A1:

Algorithm	Constraint
Hard C-Means	$\forall i,j{:}u_{i,j} \in \{0,1\}, \sum_{j'=1}^{NC} u_{i,j'} = 1$
Fuzzy C-Means[40][41]	$\forall i,j{:}u_{i,j} \in [0,1], \sum_{j'=1}^{NC} u_{i,j'}^{p} = 1, p \in [0,1]$
Possibilistic C-Means[42][43]	$\forall i,j{:}u_{i,j} \in [0,1], \sum_{i'=1}^{NS} \{1-u_{i,j}^{p}\}^{1/p} = 1, p \in [0,1]$

Table A1. Constraint on $\boldsymbol{U}$ exerted by some classic clustering algorithms

Though these forms of constraint were not explicitly stated in the proposal of the corresponding algorithms, they can be obtained straightforwardly by renaming the parameters or treating the regularizer as the Lagrange multiplier that reflects extra conditions.

The robustness against noise in FCM is a derivation of the constraint $\sum_{j=1}^{Nc} u_{i,j}^{p} = 1$, which reduces the attraction of ambiguous points to centroids as illustrated in Fig.A1, where we consider a $N_C = 2$ case. If one data is as likely to belong to cluster

one as to cluster two, then its total impact $u_{i,1} + u_{i,2}$is reduced to $2^{1-1/p} \leq 1$.

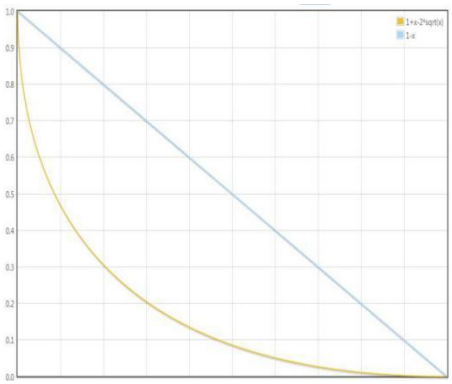


Fig.A1 The potential value of **u** in FCM, $C = 2$, $p = 0.5$.

But this setting makes FCM perform poorly when there appear to be severe overlapping in data. To relax the constraint to $\forall i,j{:}u_{i,j} \in [0,1]$ would provide the most expressive ability, however, this results in a trivial optimum $\boldsymbol{U} = \boldsymbol{0}$ for Eqn.A1. This reasoning gave birth to the constraint exerted by PCM.

However, PCM suffers from its arbitrariness in constraint form and its over-sensitivity to initialization, and there have been various studies on this topic [44][45][46] From a perspective of statistics learning, this pathology is a form of overfitting, which should be able to be compensated by a Bayesian approach.

The reason why model-based clustering algorithms instead of density [47] or kernel [48][49] based ones are adopted is that only model-based ones are capable of yielding soft result, which is more suitable to handle uncertainty.

Appendix II: A Note on Eqn.5

We show that $f(\boldsymbol{C})$ in Eqn.5 is relatively trivial compared with the first term $\sum_{i=1}^{N}\sum_{j=1}^{M} u_{i,j} \cdot d(\boldsymbol{x}_i,\boldsymbol{c}_j)$ when a gradient-based optimization of $\boldsymbol{C}$ is adopted.

The exact form of $f(\boldsymbol{C})$ can be formulated by combing Eqn.2 and Eqn.3 with Eqn.5:

$$f(\boldsymbol{C}) = \sum_{i=1}^{N_S} \ln Z(\boldsymbol{u_i},\boldsymbol{C}) \tag{A2}$$

Taking gradient of $-\ln p\,(\boldsymbol{X}|\boldsymbol{U},\boldsymbol{C})$ w.r.t $\boldsymbol{c}_j$ yields:

$$\frac{\partial}{\partial \boldsymbol{c}_j}\{-\ln p\,(\boldsymbol{X}|\boldsymbol{U},\boldsymbol{C})\} = \sum_{i=1}^{N_s} u_{i,j} \cdot \frac{\partial}{\partial \boldsymbol{c}_j} d(\boldsymbol{x}_i,\boldsymbol{c}_j) + \frac{\partial}{\partial \boldsymbol{c}_j} f(\boldsymbol{C})$$

$$= \sum_{i=1}^{N_s} u_{i,j} \cdot \frac{\partial}{\partial \boldsymbol{c}_j} d(\boldsymbol{x}_i, \boldsymbol{c}_j) + \sum_{i=1}^{N_s} \frac{1}{Z(\boldsymbol{u_i}, \boldsymbol{C})} \frac{\partial}{\partial \boldsymbol{c}_j} Z(\boldsymbol{u_i}, \boldsymbol{C}) \qquad \text{(A2)}$$

$$= \sum_{i=1}^{N_s} u_{i,j} \cdot \{ \frac{\partial}{\partial \boldsymbol{c}_j} d(\boldsymbol{x}_i, \boldsymbol{c}_j) - \int p(\boldsymbol{x}|\boldsymbol{u_i}, \boldsymbol{C}) \cdot \frac{\partial}{\partial \boldsymbol{c}_j} d(\boldsymbol{x}, \boldsymbol{c}_j) \, \mathrm{d}\boldsymbol{x} \}$$

It turns out that the gradient term $\frac{\partial}{\partial \boldsymbol{c}_j} d(\boldsymbol{x}, \boldsymbol{c}_j)$ introduced by $f(\boldsymbol{C})$ have coefficient $p(\boldsymbol{x}|\boldsymbol{u_i}, \boldsymbol{C})$. In the ideal setting, for all i, the conditional probability $p(\boldsymbol{x}|\boldsymbol{u_i}, \boldsymbol{C})$ should have its model at $\boldsymbol{x}_i$, leaving the integrand $p(\boldsymbol{x}|\boldsymbol{u_i}, \boldsymbol{C}) \cdot \frac{\partial}{\partial \boldsymbol{c}_j} d(\boldsymbol{x}, \boldsymbol{c}_j)$ to be zero at $\boldsymbol{x} \neq \boldsymbol{x}_i$. Thus, the integral $\int p(\boldsymbol{x}|\boldsymbol{u_i}, \boldsymbol{C}) \cdot \frac{\partial}{\partial \boldsymbol{c}_j} d(\boldsymbol{x}, \boldsymbol{c}_j) \, \mathrm{d}\boldsymbol{x}$ can be reduced to $\text{Constant} \cdot \frac{\partial}{\partial \boldsymbol{c}_j} d(\boldsymbol{x}_i, \boldsymbol{c}_j)$. Summarizing over $i = 1$ to N_s yields the result of Eqn.A2 as $(1 - \text{Constant}) \sum_{i=1}^{N_s} u_{i,j} \cdot \frac{\partial}{\partial \boldsymbol{c}_j} d(\boldsymbol{x}_i, \boldsymbol{c}_j)$ which makes no significant difference when we are to set it to zero. This is tantamount to dropping $f(\boldsymbol{C})$ and expect that the dominating terms in the first term are sufficient to yield the optimal $\boldsymbol{C}$.