Applied Soft Computing 20 (2014) 4–14

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection

Chien-Hsing Chen *

Department of Information Management, Ling Tung University, Taiwan

ARTICLE INFO

Article history:
Received 29 January 2013
Received in revised form 30 July 2013
Accepted 15 October 2013
Available online 1 December 2013

Keywords:
Breast cancer diagnoses
Feature selection
Cluster analysis
Filter model
Wrapper model

ABSTRACT

Models based on data mining and machine learning techniques have been developed to detect the disease early or assist in clinical breast cancer diagnoses. Feature selection is commonly applied to improve the performance of models. There are numerous studies on feature selection in the literature, and most of the studies focus on feature selection in supervised learning. When class labels are absent, feature selection methods in unsupervised learning are required. However, there are few studies on these methods in the literature. Our paper aims to present a hybrid intelligence model that uses the cluster analysis techniques with feature selection for analyzing clinical breast cancer diagnoses. Our model provides an option of selecting a subset of salient features for performing clustering and comprehensively considers the use of most existing models that use all the features to perform clustering. In particular, we study the methods by selecting salient features to identify clusters using a comparison of coincident quantitative measurements. When applied to benchmark breast cancer datasets, experimental results indicate that our method outperforms several benchmark filter- and wrapper-based methods in selecting features used to discover natural clusters, maximizing the between-cluster scatter and minimizing the within-cluster scatter toward a satisfactory clustering quality.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Data mining and machine learning techniques have been used to analyze breast cancer diagnoses, and they have been used to create models to detect the disease early or assist the diagnosis understanding. Because many features are noisy and redundant, especially in high-dimensional data representations, the created models usually suffer from noisy features participating in the training process and then compromise a satisfactory performance. For example, in classification (or clustering) learning algorithms, biased classifiers (or clusters) obtained using noisy and redundant features in the forecasting (or partitioning) process are not reliable. In the literature, many studies on feature selection are found in the literature [1–4], and the methods are grouped for supervised or unsupervised learning. In supervised learning, a feature selection method usually selects a subset of features to build a classifier that is expected to perform better than other classifiers which are built using all of the features. For example, Chhatwal et al. [5] reported that a logistic regression model can discriminate difference between benign and malignant tumors when breast cancer is detected early, and the model can identify the most important features associated with breast cancer.

However, when class labels are absent during training, supervised learning methods would be ineffective for analyzing breast cancer diagnoses. In general, a labeled breast cancer diagnosis must be labeled by an expert in the breast cancer domain, and thus, it is difficult and time-consuming to obtain, especially for new and clinical diagnoses. Furthermore, a testing record, e.g., with a cell mutation, would be less informative if it was different from training records. Most existing models assume that the class labels are well-harvested and are available for training features. However, when the class labels are absent during training, the unsupervised learning for the feature selection work are thus integral, but rarely studied in the literature.

Our paper aims to present a hybrid intelligent model that uses the cluster analysis techniques with feature selection for analyzing clinical breast cancer diagnoses. Our model provides an option of selecting a subset of salient features (i.e., attributes) for performing clustering and comprehensively considers the use of most existing models that use all the features to perform clustering. In particular, we highlight three qualitative principles that help users, e.g., clinical doctors, to analyze clinical breast cancer diagnoses effectively. First, clusters built by a subset of salient features are more practical and interpretable than clusters built using all of the features, which include noise. Because most existing studies perform clustering using features that are noisy and redundant, the clustering results are usually difficult to understand and interpret, especially for high-dimensional data. Second, clustering results

* Tel.: +886 912311895.
E-mail address: ktfive@gmail.com

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.10.024

provide clinical doctors with an understanding of the context of breast cancer diagnoses. Similarities and differences between records within a cluster or among clusters are transparent because partitioning the dataset into clusters maximizes the between-cluster scatter and minimizes the within-cluster scatter [6]. The similarities and differences then help to diagnose whether a patient suffers from breast cancer. For example, SOM-based (self-organizing map) approaches [7] show the relationships between neurons (nodes) on the map. This visual representation allows clinical doctors to observe that the breast cancer diagnoses projected for near-neighbor nodes are similar to one another but different from nodes that are farther away on the map. Finally, an efficient search for relevant records can be performed when clusters are obtained without noisy features. If a doctor wants to find similar breast cancer diagnoses from a past history dataset, a search can start at a satisfactory cluster and be expanded to neighboring clusters [8]. Because the noisy features are ignored, the time complexity of the search is significantly reduced. These three principles rely on the use of salient features to discover natural clusters using clustering learning algorithms and are applicable only to unsupervised learning.

Specifically, our presented model has three major components. The first component includes the implementation of our previous method [9] and many benchmark feature selection methods to assess their performance in the analysis of two breast cancer diagnosis datasets. Second, we use some well-known clustering learning algorithms to partition data into clusters using the identified subsets of features. The third component works on interpreting the cluster results and reporting the findings from the cluster analysis. For the purpose of performance comparison on coincident quantitative measurements, we analyze performance of our method compared to other methods in performing clustering using the identified subsets of features.

The paper is organized as follows. Section 2 reviews studies related to filter- and wrapper-based feature selection methods in supervised and unsupervised learning. Section 3 explains our presented model. Section 4 gives the experimental results. A brief discussion is given in Section 5. Finally, Section 6 details the conclusions.

2. Background studies for feature selection

For the last decade, feature selection has been prominent in the literature. This section gives a brief review of basic concepts of prior studies that relate to the development of feature selection methods in selecting salient features for the cluster analysis.

In the literature, feature selection methods are generally grouped into two categories: filter [1] and wrapper [2] models. The wrapper model depends on the classification or clustering algorithm, whereas the filter model is independent of such algorithms. The purpose of filter- and wrapper-based feature selection methods for selecting a subset of salient features is to maximize an objective criterion. The functions as the criteria include saliency, scatter separability, entropy, smoothness, consensus, density and reliability. The mutual information criterion is one of the most robust and highest-order statistic to identify salient features [10]. With respect to consider that our presented model aims to use the cluster analysis techniques for analyzing breast cancer diagnoses, we briefly introduce some unsupervised filter- and wrapper- methods, which are available for the experiments.

2.1. Filter model based on unsupervised learning

Several feature extraction techniques in unsupervised learning have been considered for dimensionality reduction. The methods identify a mapping from a high-dimensional space to a low-dimensional space with a minimal loss of information. Most existing technologies use the variance to identify the mapping. Principal component analysis (PCA) is one of the most well-known techniques for dimensionality reduction. It maps data in a high-dimensional data space to a lower-dimensional one with topological preservation. The similarity and relationship among data in the mapped space are then better understood rather than that in the original representation space. For finding this mapping, the core in PCA is a paradigm to maximize the variance. Intuitively, a larger variance of a random variable raises the between-cluster scatter [11,12]. The opposite example for this scatter is the uniform distribution of another random variable.

The above review stimulates us study the variance (Var.) metric for the feature selection process because many feature extraction techniques are based on the use of the variance metric. We also use the Max-Relevance (Max-Rel), which was implemented in [13] and is based on mutual information. In addition, we study PCA, which is commonly used to represent data in experiments.

2.2. Wrapper model based on unsupervised learning

To select features to discover natural clusters, wrapper-based methods in unsupervised learning require a pre-specified clustering learning algorithm to partition a dataset into subsets (clusters). Once the clustering process converges (e.g., the optimal clusters are obtained), features are learned from the clusters [14,15]. Nonparametric methods for wrapper feature selection have been widely developed in the literature [16,17]. Additionally, several studies used the internal or the relative cluster validity as the criterion to observe how a feature was informative to the clusters, when a dataset was partitioned into clusters. For example, Chow et al. [18] proposed a feature selection method based on the compactness of and separation from clusters. Huang et al. [19] proposed a feature co-selection for Web document clustering that clusters the results in one type of feature space to identify salient features in other types of feature spaces. Li et al. [20] proposed a new text clustering method with feature selection and extended the chi-square term-category independence test to measure whether the dependency between a term and a cluster was positive or negative. Sanguinetti [21] presented a latent variable model to perform dimensional reduction on a dataset that contained clusters. In his study, a variable was salient if it preserved clustered information from the original representation space when mapped to a latent space.

We follow the literature studies and implement some wrapper-based methods that learn features from the clusters resulting from a particular clustering algorithm, e.g., *A*. "wrapper-*A*" represents a wrapper-based feature selection method. We use *A* to obtain $M \in \{2, 3, \ldots, 30\}$ clusters for a dataset. We use the Max-Dependency criterion [13], based on mutual information, to evaluate the relevance of a feature to its labeled cluster and select salient features according to their relevance. Four clustering algorithms, *K*-means, self-organizing map (SOM), complete-link hierarchical clustering (HC) and partitioning around medoids (PAM), are applied for the wrapper model. We thus have four methods including wrapper-Kmeans, wrapper-SOM, wrapper-HC and wrapper-PAM. Because similar units can be grouped in SOM, we follow a previous study [22] to obtain the user-defined number of clusters, *M*, for the map units.

3. Our hybrid intelligent model

We present a hybrid intelligent model that uses the cluster analysis techniques with feature selection for analyzing clinical

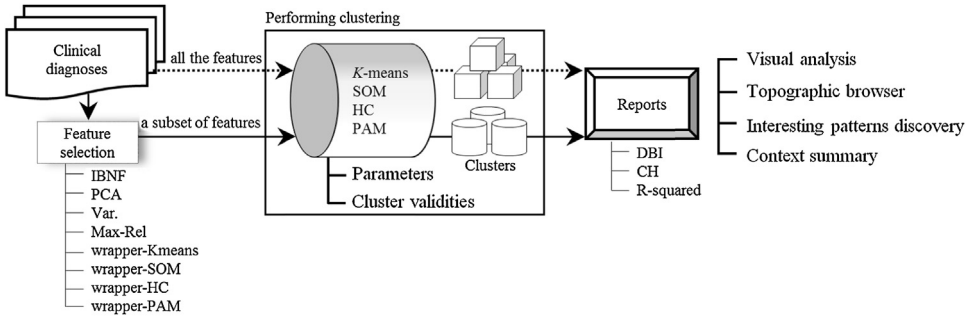

Fig. 1. A hybrid intelligent model for visual clustering with feature selection for the analysis of breast cancer diagnoses.

breast cancer diagnoses. Our model provides an option of selecting a subset of salient features for performing clustering and comprehensively considers the use of most existing models that use all the features to perform clustering. This model starts with inputting diagnoses. Before performing clustering using clustering learning algorithms, we apply several feature selection methods, as briefly discussed in Section 2, to select salient features. The model is introduced in Fig. 1. The solid lines represent our clustering learning process upon using our presented feature selection method, whereas the dotted lines represent the existing clustering learning process. Considering the use of both processes provides a satisfactory comparison of coincident quantitative measurements for analyzing breast cancer diagnoses. Once the clusters are partitioned, we have reports with respect to visual analysis, topographic browser, interesting patters discovery and context summary. For example, SOM-based methods are very popular for yielding the clusters in a visualization format, resulting satisfactory reports.

3.1. Our feature selection method

We utilize our proposed feature selection method, which was an Instance-Based learning to quantify features from the Nearest and Farthest neighbors (which we call IBNF [9]), to adaptively analyze breast cancer diagnoses. Consider a clustering characteristic in that (1) a data instance and its nearest neighbors are usually clustered in a cluster, and (2) this data instance and its farthest neighbors are usually clustered in different clusters. IBNF was then motivated by this clustering characteristic. In particular, a feature might be more salient if this feature helps to minimize distances between each instance and its nearest neighbors, and it helps to maximize distances between each instance and its farthest neighbors.

IBNF denotes a dataset $\boldsymbol{X}$ consisting of n data instances $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = [x_{1,i}, \ldots, x_{j,i}, \ldots, x_{d,i}]^T$ means the ith instance in $\boldsymbol{X}$ with d dimensions. IBNF also defines a non-zero feature vector $\mathbf{w}(t) = [w_1(t), \ldots, w_j(t), \ldots, w_d(t)]^T$, where the element $w_j(t)$ is a real-valued quantity at the tth iteration. The $\mathbf{w}(t)$ reflects the "weights" for the weighted Euclidean distance metric applied to calculate distances between data instances and is utilized to obtain the nearest and farthest neighbors. Each $\mathbf{x}_i$ has its nearest and farthest neighbors while inputting $\mathbf{w}(t)$. IBNF then evaluates feature salience with respects these neighbors for $\mathbf{x}_i$ and outputs a feature salience vector $[\mathbf{u}_i^{\mathbf{w}(t)} = u_{1,i}^{\mathbf{w}(t)}, \ldots, u_{j,i}^{\mathbf{w}(t)}, \ldots, u_{d,i}^{\mathbf{w}(t)}]^T$ recording feature salience for $\mathbf{x}_i$ under $\mathbf{w}(t)$. In particular, the criterion for evaluating $u_{j,i}^{\mathbf{w}(t)}$ is defined using feature compactness and separability as follows (1).

$$u_{j,i}^{\mathbf{w}(t)} = FS\left(j; \mathbf{x}^{\phi}_{i\to l;\mathbf{w}(t)}\right) + FC\left(j; \mathbf{x}^{\Theta}_{i\to k;\mathbf{w}(t)}\right), \quad \text{for } k = 1, \ldots, K \text{ and } l = 1, \ldots, L \tag{1}$$

where $FS()$ evaluates feature separability and $FC()$ evaluates feature compactness for the jth feature of $\mathbf{x}_i$ under $\mathbf{w}(t)$. The $\mathbf{x}^{\Theta}_{i\to k;\mathbf{w}(t)}$ and $\mathbf{x}^{\phi}_{i\to l;\mathbf{w}(t)}$ respectively represent the nearest and the farthest neighbor, where $\mathbf{x}^{\Theta}_{i\to k;\mathbf{w}(t)} = [x^{\Theta}_{1,i\to k;\mathbf{w}(t)}, \ldots, x^{\Theta}_{j,i\to k;\mathbf{w}(t)}, \ldots, x^{\Theta}_{d,i\to k;\mathbf{w}(t)}]^T$ and $\mathbf{x}^{\phi}_{i\to l;\mathbf{w}(t)} = [x^{\phi}_{1,i\to l;\mathbf{w}(t)}, \ldots, x^{\phi}_{j,i\to l;\mathbf{w}(t)}, \ldots, x^{\phi}_{d,i\to l;\mathbf{w}(t)}]^T$. When the neighbors are obtained, we evaluate the salience for each feature using $FS()$ and $FC()$, as expressed in (2) and (3).

$$FS\left(j; \mathbf{x}^{\phi}_{i\to l;\mathbf{w}(t)}\right) = \frac{1}{L}\sum_{l=1}^{L} \varepsilon\left(x_{j,i}, x^{\phi}_{j,i\to l;\mathbf{w}(t)}\right) \tag{2}$$

$$FC\left(j; \mathbf{x}^{\Theta}_{i\to k;w(t)}\right) = -\frac{1}{K}\sum_{k=1}^{K} \varepsilon\left(x_{j,i}, x^{\Theta}_{j,i\to k;w(t)}\right) \tag{3}$$

where $\varepsilon()$ is an absolute operation (e.g., $\varepsilon([0.5, 0.2]^T, [0.1, 0.3]^T) = [0.4, 0.1]^T$). The larger value of $u_{j,i}^{\mathbf{w}(t)}$ indicates that the jth feature of $\mathbf{x}_i$ under $\mathbf{w}(t)$ is more salient, and the saliences of all the features are recorded in $u_i^{\mathbf{w}(t)}$.

Because different $\mathbf{w}$ values inconsistently contribute salience $\mathbf{u}$, we should locate the best $\mathbf{u}_i^{\mathbf{w}(t)}$ by deriving $\mathbf{w}(t)$ such that every instance consistently favors the salient features. If $\{\mathbf{u}_i^{\mathbf{w}(t)}\}_{i=1}^{n}$ were consistent for every instance, we would say $\mathbf{w}(t)$ is the best solution, and the features can be selected according the elements of $\mathbf{w}$. The method to reduce the searching problem is to optimize the sum-of-squared error (i.e., $||\mathbf{u}_i^{\mathbf{w}(t)} - \mathbf{w}(t)||^2$). Specifically, we make a decision of stopping criterion to obtain the best $\mathbf{w}(T)$ which is a local optimal solution. The procedure to output the best $\mathbf{w}(T)$ for the use of analyzing diagnoses is shown in Fig. 2.

3.2. Performing clustering with feature selection

In this section, we consider the selected features for executing clustering learning algorithms. Because all the clustering learning algorithms depend on a distance metric which is used to calculate distances between data instances, we provide a justification on how the selected features are used for the distance metric adaptively used in a clustering learning algorithm (e.g., SOM).

We illustrate a system with kernel based on SOM [23] with the use of the selected salient features (i.e., using $\mathbf{w}(T)$). The SOM approach nonlinearly projects high-dimensional data onto a low-dimensional grid. The projection preserves the topological order from the input space; hence, similar data patterns in the input space would be assigned to the same map node or to nearby nodes on the trained map. Denote that we have an input pattern $\mathbf{x}_i = [x_{1,i}, \ldots, x_{j,i}, \ldots, x_{d,i}]^T$ in the input space, where $x_{j,i}$ is the value of the jth attribute of $\mathbf{x}_i$; also assume a map unit $\mathbf{m}_k = [m_{1,k}, \ldots, m_{j,k}, \ldots, m_{d,k}]^T$, where $m_{j,k}$ is the value of the jth

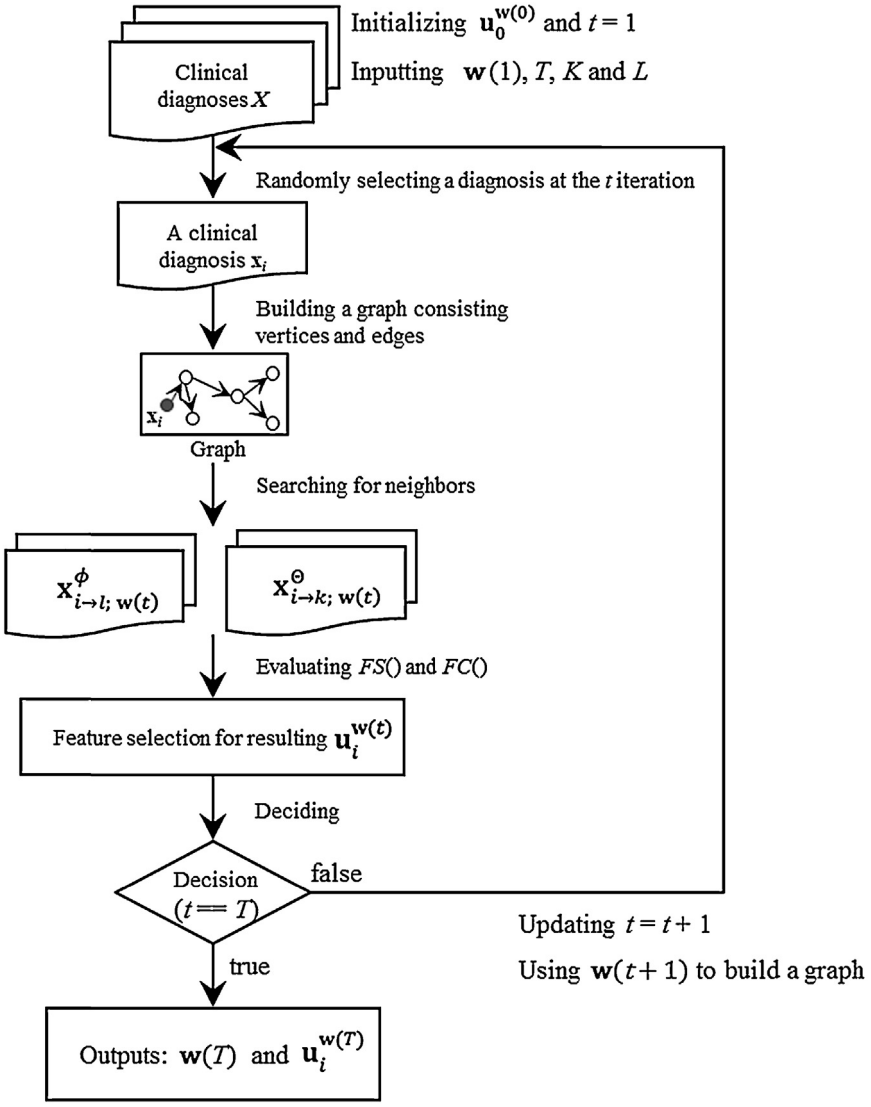

Fig. 2. Procedure of harvesting $\mathbf{w}(T)$ and $\mathbf{u}_i^{\mathbf{w}(T)}$ for the use of selecting salient features for clinical diagnoses, where the features are selected according to the elements of $\mathbf{w}(T)$.

Table 1
Implementation of feature selection methods, clustering learning algorithms and cluster validations.

Categories	Methods	Descriptions
Feature selection methods	IBNF	Procedure of IBNF is present in Fig. 2.
	PCA	Use the eigenvector of the first principal component
	Var.	Use variance statistic to evaluate features
	Max-Rel	[13]
	Wrapper-Kmeans	Background study in Section 2.2.
	Wrapper-SOM	Background study in Section 2.2.
	Wrapper-HC	Background study in Section 2.2.
	Wrapper-PAM	Background study in Section 2.2.
Clustering learning algorithms	*K*-means	[24]
	SOM	[25]
	HC	[24]
	PAM	[26]
Cluster validations	DBI	[27]
	CH	[27]
	R-squared	[28]

attribute of the kth unit in the map. This system therefore has two major processes. The first process is to search for the best matching unit (BMU) from all of the map units for each input pattern, where the BMU is most similar to the input pattern. The method to search for this BMU thus requires a distance metric to compute distances between input patters and map units. The Euclidean metric is an option; specifically, we consider the outputs of the selected features by exploring $\mathbf{w}(T)$ for computing distance between $\mathbf{x}_i$ and $\mathbf{m}_k$, namely, $dis(\mathbf{x}_i, \mathbf{m}_k|\mathbf{w}(T))$ written as follows.

$$dis(\mathbf{x}_i, \mathbf{m}_k|\mathbf{w}(T)) = \left(\sum_{j=1}^{d} \left((x_{j,i} - m_{j,k}) \times w_j(T) \right)^2 \right)^{1/2} \quad (4)$$

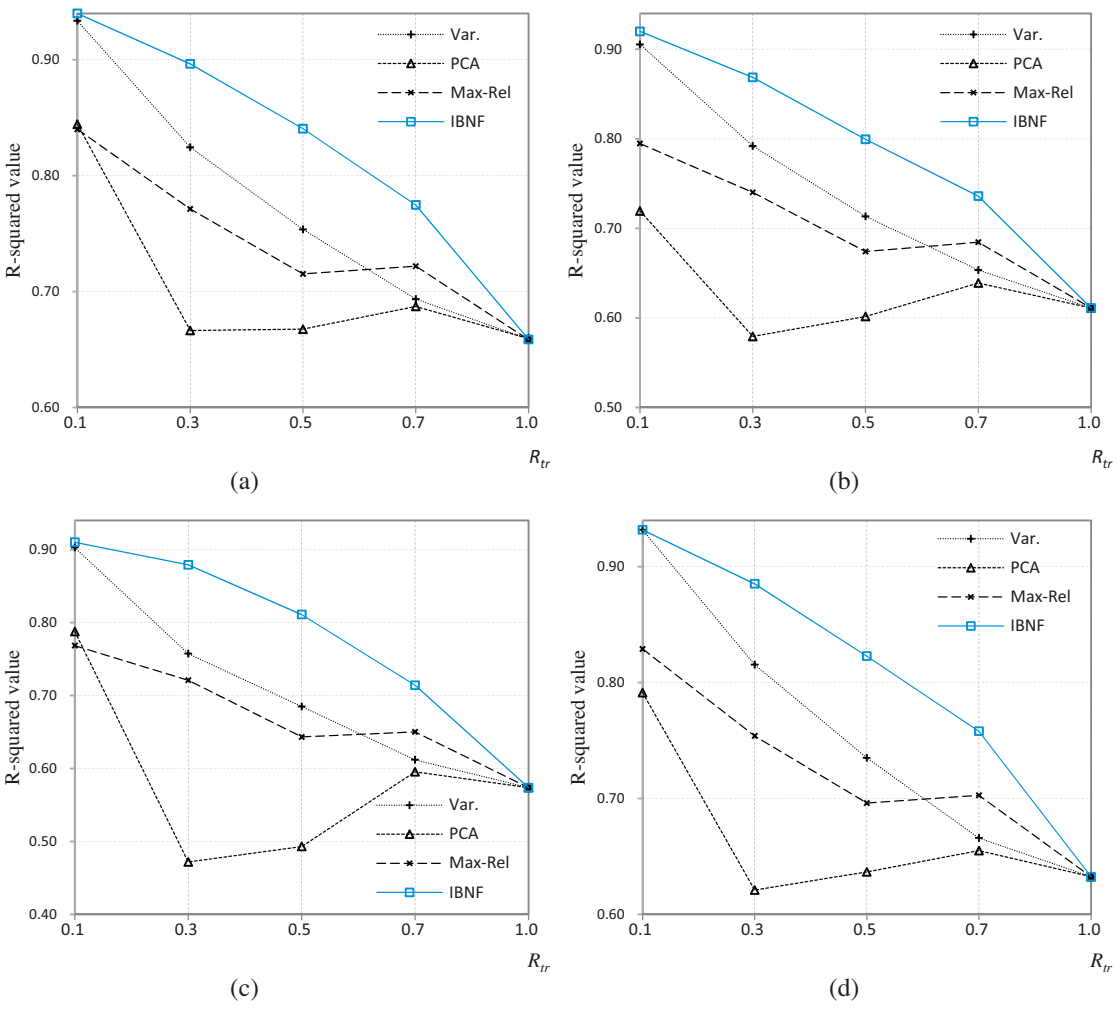

Fig. 3. Comparison of *R*-squared measurements for various R_{tr} on the Breast cancer Wisconsin (Diagnostic) dataset. For each feature selection method, the selected features are employed to perform clustering using (a) *K*-means, (b) SOM, (c) HC and (d) PAM.

The second process is to update the weights onto the map units, where we follow the study [23] for this update. When inputting an input pattern $\mathbf{x}_i$ for training SOM, the update function for a neighborhood node $\mathbf{m}_k$ is written as follows.

$$\mathbf{m}_k(t+1) = \mathbf{m}_k(t) + \alpha(t) \times h_k(t) \times [\mathbf{x}_i(t) - \mathbf{m}_k(t)] \quad (5)$$

where $0 < \alpha(t) < 1$ is the learning-rate function, $h_k(t)$ is the neighborhood function, which calculates the lattice distance between the BMU and $\mathbf{m}_k$, and *dis*() is the weighted Euclidean distance. Both $\alpha(t)$ and the width of $h_k(t)$ decrease gradually with increasing step *t*.

3.3. Implementations

Our model has three major components according to feature selection method, clustering learning algorithm and interpretation of cluster analysis. To validate performance comparison on coincident quantitative measurements, we implement methods, algorithms and criteria for the experiments using Matlab with version R2013a. These feature selection methods, clustering learning algorithms and cluster validations are summarized in Table 1 and are briefly introduced later.

- Implementation of feature selection methods:

We design four filter-based feature selection methods: IBNF, PCA, Var. and Max-Rel. Furthermore, we implement four wrapper-based methods: wrapper-Kmeans, wrapper-SOM, wrapper-HC and wrapper-PAM. These eight methods are briefly introduced in Section 2.

- Implementation of clustering learning algorithms

  To demonstrate performance of the selected features in identifying natural clusters, the selected features are tested in some well-known clustering learning algorithms to partition a dataset into clusters. In particular, we use *K*-means, SOM, HC and PAM algorithms, which are available in the study [29], to perform clustering. We followed a previous study [22] to obtain the user-defined number of clusters for the map units because similar units can be grouped in SOM.
- Implementation of cluster validations:

  To evaluate clustering performance, three cluster validations are implemented: the Davies-Bouldin index (DBI), the Calinski-Harabasz (CH) index [27] and the *R*-squared validity (*R*-squared) [28]. The DBI is chosen because it is popular in cluster analysis. A lower DBI value indicates higher compactness within a cluster or higher separability between clusters. We also use the CH index to measure between-cluster isolation and within-cluster coherence. A larger CH value represents better performance. Furthermore, we adaptively use *R*-squared validity, which has

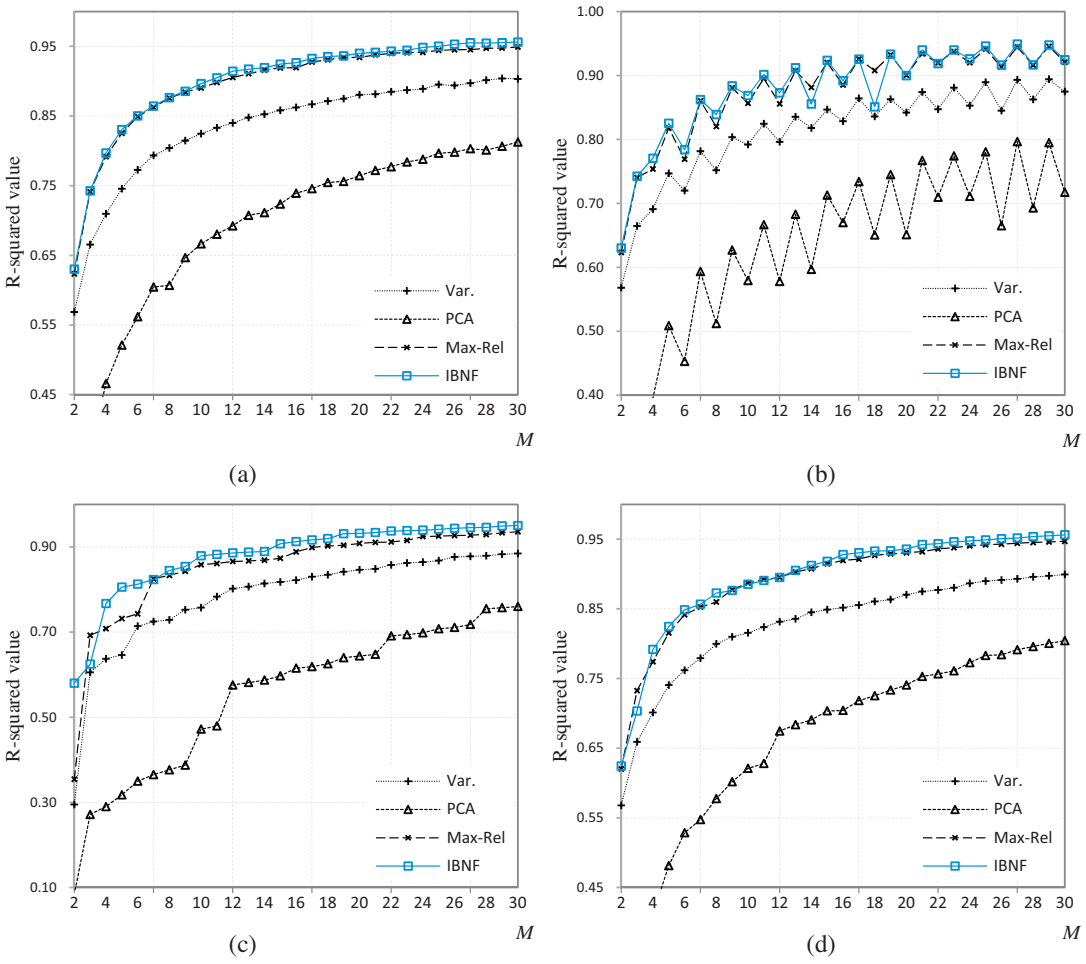

Fig. 4. Comparison of R-squared measurements for several values of M on the Breast cancer Wisconsin (Diagnostic) dataset. The ratio, R_{tr}, is 0.3, and the selected features are employed to perform clustering using (a) K-means, (b) SOM, (c) HC and (d) PAM.

been used to measure interpretation capability in the statistical model.

4. Experiments

Extensive experiments have been conducted to evaluate our method against four filter- and four wrapper-based feature selection methods in the context of clustering breast cancer diagnoses. The experimental setting to observe performance and validation was shown as follows.

- **Defined a suitable size of the selected features**: When a subset of salient features was selected using each of these methods, the selected subset of features was applied for performing clustering. We set the ratio to determine the number of selected features, $R_{tr} \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$. For example, if $R_{tr} = 0.3$, the number of the selected features was $d \times 0.3$; if $R_{tr} = 1.0$, all features were used.
- **To partition diagnoses into clusters**: We performed clustering using the selected subset of features. Because clustering was usually an ill-posed problem implicating that some assumptions should be made in advance, we should not know the correct number of clusters. Therefore, we expected to have $M \in \{2, 3, \ldots, 30\}$ clusters resulting from each of the clustering learning algorithms. We then observed the cluster quality of varying number of clusters.

4.1. Performance results of the filter model

The first set of experiments utilized our IBNF method and compared it to some existing filter-based clustering feature selection methods. We performed the experiments on the Breast cancer Wisconsin (Diagnostic) data set [30], available from the UCI Machine Learning Repository.[1] The number of instances is 569, and the number of features is 32. This dataset consists of an ID number and a class label, where each data instance is labeled either malignant or benign. The other 30 features describe characteristics of the cell nuclei for each image digitized from a breast mass. These features are characterized from the descriptions including radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, concave points, and fractal dimension. The method to obtain these features was referred to the study [31].

We briefly observed the salient features selected using IBNF and Var., where the features were ordered according to their saliences. In our particular use, we ignored two features: ID number and class label. Thus, only 30 features were retained. The comparison result is shown in Table 2. The second and the third columns represent the ordered features. For example, the most salient feature for IBNF is the fourth feature; for Var., the 28th feature. We see that IBNF and Var. output the salient features differently. IBNF favors the features described from the radius and the texture of cell nuclei,

[1] http://www.ics.uci.edu/~mlearn/MLRepository.html

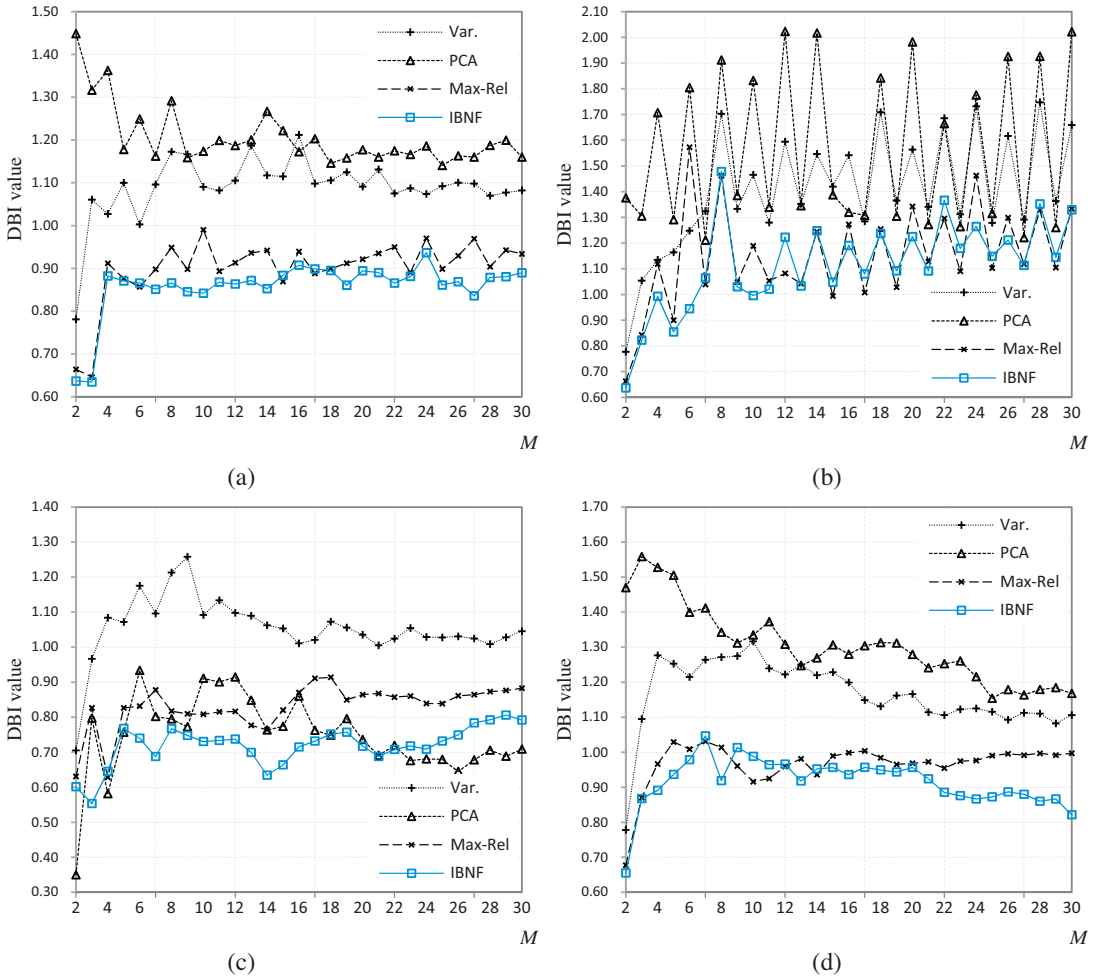

Fig. 5. Comparison of DBI measures for several values of M on the Breast cancer Wisconsin (Diagnostic) dataset. The ratio, R_{tr}, is 0.3, and the selected features are employed to perform clustering using (a) K-means, (b) SOM, (c) HC and (d) PAM.

Table 2
Comparison of the salient features selected using IBNF and Var.

Order	IBNF	Var.
1	4	28
2	1	8
3	3	7
4	21	21
5	24	3
.	.	.
.	.	.
.	.	.
26	10	11
27	29	13
28	9	20
29	30	14
30	20	17

whereas Var. favors those described from the concave points and perimeter.

We then tested whether the salient features are effective for performing clustering. Specifically, four feature selection methods were utilized to select features, which were then applied to perform clustering using K-means, SOM, HC and PAM. Furthermore, we further observed how the varying sizes of the selected features contribute the partitioned clusters. To evaluate the quality of the partitioned clusters, we used R-squared validation to evaluate the performance of each method in partitioning the dataset into clusters using different sizes of the selected features. The results of this evaluation for several values of R_{tr} are shown in Fig. 3.

Fig. 3 shows that clustering obtained using a subset of features is better than that obtained using all of the features. For most feature selection methods, the features are well ordered by their salience. The R-squared value is usually largest at $r = 0.1$ and decreases through $r = 1$. This indicates that a lower number of salient features, i.e., $r = 0.1$, contribute force to clustering. In addition, our method performs better than the other methods in selecting feature subsets for performing clustering, and PCA is often ineffective for selecting features for discovering clusters. This experiment indicates that many features are noisy and that feature selection is demanding, especially for high-dimensional data.

Next, we observed how the selected features affected the obtained clusters. In particular, we analyzed the performance in partitioning dataset into various clusters $M \in \{2, 3, \ldots, 30\}$. The ratio, R_{tr}, was set to 0.3 to optimize the performance for most methods because this setting led the R-squared value improved, especially the comparison to PCA. The results of this evaluation for various M are shown in Fig. 4. In Fig. 4, we see that IBNF often performs better than the other methods in selecting features for performing clustering. The performances of IBNF and Max-Rel are comparable, and PCA performs poorly. Recall that the R-squared validation has been commonly used to measure how an independent variable can predict a dependent one. In this experiment, we see that the selected features can be used to interpret the obtained clusters.

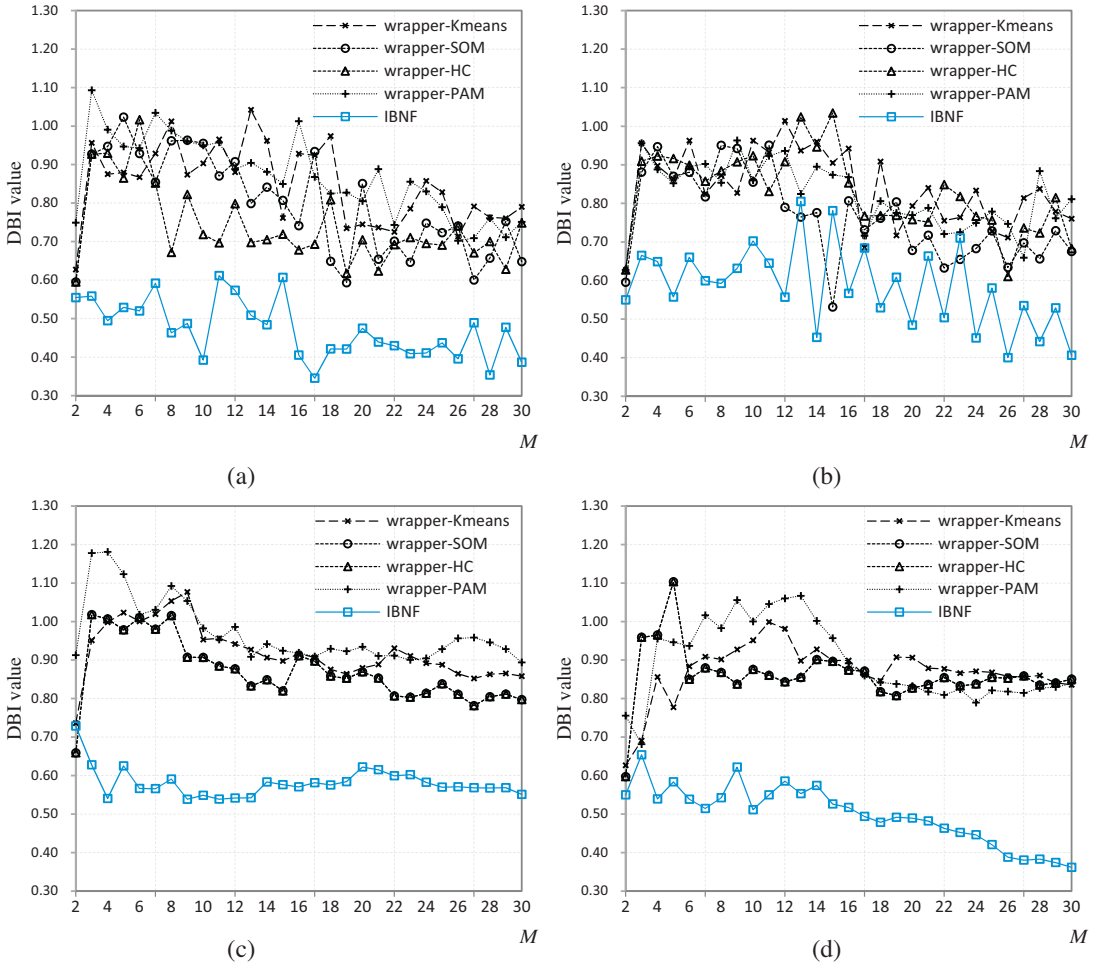

Fig. 6. Comparison of DBI measures for several values of M on the Breast cancer Wisconsin (Original) dataset. The ratio, R_{tr}, is 0.5, and the selected features are employed to perform clustering using (a) K-means, (b) SOM, (c) HC and (d) PAM.

The purpose of clustering is to maximize the between-cluster scatter and minimize the within-cluster scatter. We used the DBI to measure performance in terms of between- and within-cluster scatters. The ratio, R_{tr}, was set to 0.3, and the selected features were utilized to obtain clusters, $M \in \{2, 3, \ldots, 30\}$. The results are shown in Fig. 5. Often, IBNF performs better than the other methods, and PCA are ineffective for identifying clusters. In some cases, IBNF does not have the best performance. In Fig. 5(c), IBNF is ineffective for performing clustering when $M > 22$. In Fig. 5(d), IBNF is ineffective for performing clustering when M is between 9 and 12. From the overall performance, a subset of salient features is fundamental to discovering clusters. The performance of IBNF is usually the best, and Max-Rel is the second best.

4.2. Performance results on the wrapper model

The second set of experiments analyzed the performance of wrapper-based methods in selecting features used in clustering. We performed the experiments on the Breast cancer Wisconsin (Original) dataset [32], available from the UCI Machine Learning Repository. This dataset is obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. There are 683 clinical cases, each of which is labeled either benign or malignant. To perform an unsupervised learning for feature selection, we ignored two attributes: ID number and class label. We thus had nine attributes including clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. Here, we briefly observed the salient features selected using IBNF and Var., where the features were ordered according to their saliences. Table 3 shows the ordered features, and we see the sixth feature (i.e., bare nuclei) is most salient for the both methods.

Table 3
Comparison of the salient features selected using IBNF and Var.

Order	IBNF	Var.
1	6	6
2	8	2
3	3	8
4	2	3
5	4	4
6	5	1
7	7	7
8	9	5
9	1	9

We further used four wrapper-based methods, namely, wrapper-Kmeans, wrapper-SOM, wrapper-HC and wrapper-PAM, for the experiments. Thus, IBNF and four wrapper-based feature selection methods were applied to select subsets of features, which were applied for performing clustering. To demonstrate the effectiveness of the selected feature subset in clustering, half of the features were selected to obtain clusters, $M \in \{2, 3, \ldots, 30\}$, using clustering learning algorithms. We first used the DBI to demonstrate performance, as shown in Fig. 6.

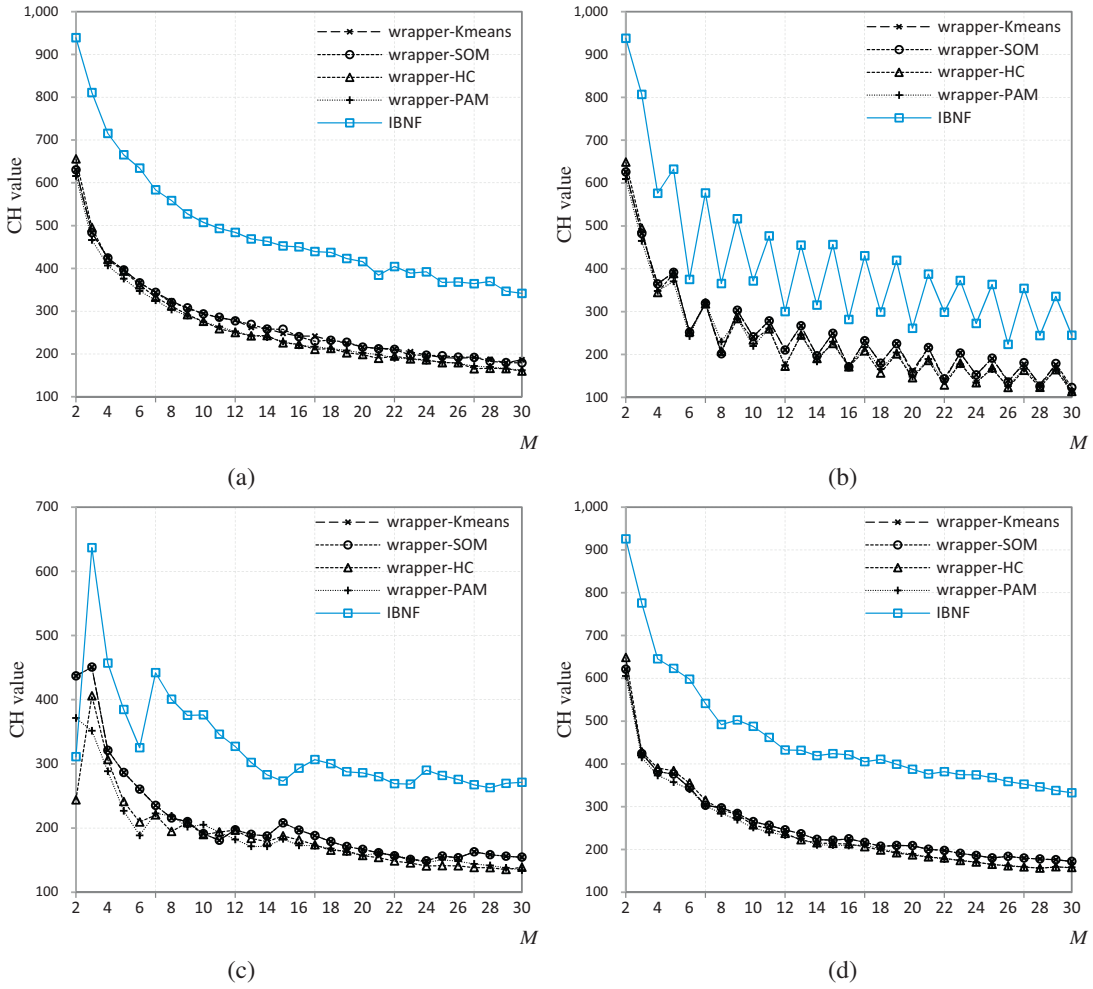

Fig. 7. Comparison of CH index measures for several values of M on the Breast cancer Wisconsin (Original) dataset. The ratio, R_{tr}, is 0.5, and the selected features are applied to perform clustering using (a) K-means, (b) SOM, (c) HC and (d) PAM.

In Fig. 6, IBNF usually outperforms the wrapper-based methods in selecting features to be used for performing clustering. In Fig. 6(b), the wrapper-SOM performs better than the other wrapper-based methods. In the feature selection process, the wrapper-SOM learns features from the clusters resulting from SOM. Thus, the selected features are effective in identifying clusters using SOM. However, most wrapper-based methods are less flexible because the selected features learned from a particular clustering algorithm cannot be adapted for use by another clustering algorithm.

We used a second cluster validation to measure the performance in selecting features used in identifying clusters. We used the CH index to evaluate performance, as shown in Fig. 7. We see that IBNF outperforms the other models in discovering clusters. However, Fig. 7(c) shows that IBNF fails when $M = 2$.

Because the purpose of clustering is to partition a dataset into clusters such that the instances are similar within a cluster and different between clusters, IBNF better identifies features. Because IBNF is based on the filter model, the features are selected regardless of any clustering learning algorithm used for training features. In contrast, wrapper-based methods use all of the features, including noise, to explore clusters. Features that are informative to the clusters are selected but suffer from noisy features that affect the cluster shape or the number of clusters. In addition, the biased clusters lead to poorly select features. To identify natural and interesting clusters, the selected features obtained using IBNF perform better than those obtained using wrapper-based methods.

Table 4
Cluster distribution resulting from SOM under different size of features.

ID no.	Cluster ID (half of features)	Cluster ID (all the features)
1	2	2
2	2	1
3	2	2
4	1	1
5	2	2
.	.	.
.	.	.
.	.	.
681	1	1
682	1	1
683	1	1

We then observed the cluster distribution resulting from SOM using the selected subset of features (i.e., R_{tr}, is 0.5) or all the features. The cluster distribution is shown in Table 4. The first column is the ID number, and the second and the third columns represent the cluster distribution resulting from using a half of features and all the features, respectively. We see the cluster distribution using a half of features is different from that using all the features. This indicates that the clustering quality relies on what the features selected. From our above experiments, the selected subset of features is helpful for the diagnoses with respect to cluster analysis toward a satisfactory clustering quality.

5. Discussion

Our paper uses the cluster analysis techniques with feature selection for analyzing breast cancer diagnoses. To demonstrate performance of many feature selection methods in clustering breast cancer diagnosis data, we used coincident quantitative measurements to analyze the salient features used to discover clusters. We implemented IBNF and three filter-based feature selection methods: PCA, Var. and Max-Rel. The selected features were utilized to perform clustering using four well-known clustering learning algorithms: *K*-means, SOM, HC and PAM. Performance was evaluated using DBI, CH and *R*-squared validities, which are commonly used in statistical model and cluster analysis. The experimental results, as discussed in Section 4.1, show that the features selected by the four feature selection methods are effective in discovering clusters, maximizing the between-cluster scatter and minimizing the within-cluster scatter, and the performance of IBNF is usually the best.

We also implemented four wrapper-based feature selection methods: wrapper-Kmeans, wrapper-SOM, wrapper-HC and wrapper-PAM. In the experiment, we studied whether the selected features that were learned from clusters produced by a particular clustering algorithm could be adapted for use by another clustering algorithm. The experimental results are discussed in Section 4.2. We see that the selected features, which are specific to a particular clustering algorithm, cannot be adapted for use by another clustering algorithm. For example, the features selected using the wrapper-Kmeans are not effective for performing clustering by the SOM, PAM or HC algorithms. This finding is interesting but less mentioned in the literature.

IBNF is motivated by a clustering characteristic: (1) a data instance and its nearest neighbors are usually clustered in a cluster, and (2) this data instance and its farthest neighbors are usually clustered in different clusters. IBNF then selects salient features if these features help to minimize distances between each instance and its nearest neighbors, and they help to maximize distances between each instance and its farthest neighbors. Therefore, comparing to other implemented filter-based feature selection methods, which do not consider any clustering characteristic or clustered information, IBNF outperforms than these filter-based feature selection methods in selecting features for discovering clusters. On the other hand, the wrapper-based feature selection methods use all the features, which might include noisy and redundant features, to obtain clusters resulting from the prespecific clustering learning algorithms. The obtained clusters are thus evil, and the selected features are biased to these clusters suffering from the participation of the noisy features. IBNF thus identifies a better subset of features to discover clusters and performs better performance than the implemented wrapper-based feature selection methods.

6. Conclusions

Feature selection is one of the most effective methods to enhance data representation and improve performance in terms of specified criteria, e.g., generalization classification accuracy. In the literature, many studies select a subset of salient features using supervised learning rather than unsupervised learning. When the class labels are absent during training, feature selection in unsupervised learning is integral, but its extensible application is rarely studied in the literature.

The objective of this study is to select salient features that can be used to identify interesting clusters in the analysis of breast cancer diagnoses. Specifically, we highlight three qualitative principles that help users to analyze clinical breast cancer diagnoses using clusters resulting from a subset of salient features. First, the clusters built by a subset of salient features are more practical and interpretable than those built by all of the features, which include noise. Second, the clustering results provide clinical doctors with an understanding of the context of clinical breast cancer diagnoses. Finally, a search for relevant records based on the clusters obtained when noisy features are ignored is more efficient. These three principles rely on the discovery of natural clusters using salient features and are applicable only to unsupervised learning.

To demonstrate the usefulness of these three qualitative principles, we use coincident quantitative measurements to analyze the salient features for discovering clusters. Eight feature selections methods, four clustering learning algorithms and three cluster validations are implemented using Matlab software. The experiments on the Breast cancer Wisconsin (Diagnostic) and Breast cancer Wisconsin (Original) datasets demonstrate that the selected features are effective for selecting salient features to discover natural clusters. Based on a performance evaluation using well-known validations in statistical model and cluster analysis, our analysis provides an interesting aspect in feature selection for discovering clusters. Our method might significantly impact data mining and machine learning applications in the future.

Acknowledgements

This research was supported by the National Science Council, Taiwan under grant NSC 101-2410-H-275-012.

References

[1] H. Liu, L. Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on Knowledge and Data Engineering 17 (4) (2005) 491–502.
[2] M.A. Jayaram, A.G. Karegowda, A.S. Manjunath, Feature subset selection problem using wrapper approach in supervised learning, International Journal of Computer Applications 1 (7) (2010) 13–16.
[3] M. Shah, M. Marchand, J. Corbeil, Feature selection with conjunctions of decision stumps and learning from microarray data, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (1) (2012) 174–186.
[4] D. Sun, D. Zhang, Bagging constraint score for feature selection with pairwise constraints, Pattern Recognition 43 (2010) 2106–2118.
[5] J. Chhatwal, O. Alagoz, M.J. Lindstrom, C.E. Kahn, K.A. Shaffer, E.S. Burnside, A logistic regression model based on the national mammography database format to aid breast cancer diagnosis, American Journal of Roentgenology 192 (4) (2009) 1117–1127.
[6] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (4) (1979) 224–227.
[7] H. Yin, Data visualization and manifold mapping using the ViSOM, Neural Networks 15 (2002) 1005–1016.
[8] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, A. Saarela, Self organization of a massive document collection, IEEE Transactions on Neural Networks 11 (2000) 574–585.
[9] C.-H. Chen, Feature selection for unlabeled data, in: Proceedings of the Second International Conference on Swarm Intelligence (LNCS 6729), Chongqing, China, 2011, pp. 269–274.
[10] T.W.S. Chow, D. Huang, Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information, IEEE Transactions on Neural Networks 16 (1) (2005) 213–224.
[11] R. Varshavsky, A. Gottlieb, M. Linial, D. Horn, Novel unsupervised feature filtering of biological data, Bioinformatics 22 (14) (2006) 507–513.
[12] M.H.C. Law, M.A.T. Figueiredo, A.K. Jain, Simultaneous feature selection and clustering using mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (9) (2004) 1154–1166.
[13] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.
[14] L.I. Kuncheva, D.P. Vetrov, Evaluation of stability of k-means cluster ensembles with respect to random initialization, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (11) (2006) 1798–1807.
[15] J.G. Dy, C.E.B.A. Kak, L.S. Broderick, A.M. Aisen, Unsupervised feature selection applied to content-based retrieval of lung images, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (3) (2003) 373–378.
[16] K.M. Carter, R. Raich, W.G. Finn, A.O. Hero, FINE: Fisher information nonparametric embedding, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (11) (2009) 2093–2098.
[17] Y. Sun, S. Todorovic, S. Goodison, Local-learning-based feature selection for high-dimensional data analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1610–1626.

[18] T.W.S. Chow, P. Wang, E.W.M. Ma, A new feature selection scheme using a data distribution factor for unsupervised nominal data, IEEE Transactions on Systems, Man, and Cybernetics—PART B: Cybernetics 38 (2) (2008) 499–509.
[19] S. Huang, Z. Chen, Y. Yu, W.Y. Ma, Multitype features coselection for Web document clustering, IEEE Transactions on Knowledge and Data Engineering 18 (4) (2006) 448–459.
[20] Y. Li, C. Luo, S.M. Chung, Text clustering with feature selection by using statistical data, IEEE Transactions on Knowledge and Data Engineering 20 (5) (2008) 641–652.
[21] G. Sanguinetti, Dimensionality reduction of clustered data sets, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (3) (2008) 535–540.
[22] J. Vesanto, E. Alhoniemi, Clustering of the self-organizing map, IEEE Transactions on Neural Networks 11 (3) (2000) 586–600.
[23] T. Kohonen, Self-Organization and Associative Memory, 3rd edition, Springer-Verlag, Berlin-Heidelberg-New York-Tokyo, 1989.
[24] J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
[25] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43 (1) (1982) 59–69.
[26] M. Laan, K. Pollard, J. Bryan, A new partitioning around medoids algorithm, Journal of Statistical Computation and Simulation 73 (8) (2003) 575–584.
[27] G. Shu, B. Zeng, Y.P. Chen, O.H. Smith, Performance assessment of kernel density clustering for gene expression profile data, Comparative and Functional Genomics 4 (3) (2003) 287–299.
[28] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Journal of Intelligent Information Systems 17 (2001) 107–145.
[29] K. Wang, B. Wang, L. Peng, CVAP: validation for cluster analyses, Data Science Journal 8 (2009) 88–93.
[30] W.H. Wolberg, O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, in: Proceedings of the National Academy of Sciences, U.S.A., 1990, pp. 9193–9196.
[31] K.P. Bennett, Decision tree construction via linear programming, in: Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, Utica, IL, 1992, pp. 97–101.
[32] O.L. Mangasarian, R. Setiono, W. Wolberg, Pattern recognition via linear programming: theory and application to medical diagnosis, in: SIAM Workshop on Optimization, 1990.