Expert Systems with Applications 41 (2014) 1476–1482

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms

Bichen Zheng, Sang Won Yoon *, Sarah S. Lam

Department of Systems Science and Industrial Engineering, State University of New York at Binghamton, Binghamton, NY 13902, United States

ARTICLE INFO

Keywords:
Data mining
K-means
Support vector machine
Cancer diagnosis

ABSTRACT

With the development of clinical technologies, different tumor features have been collected for breast cancer diagnosis. Filtering all the pertinent feature information to support the clinical disease diagnosis is a challenging and time consuming task. The objective of this research is to diagnose breast cancer based on the extracted tumor features. Feature extraction and selection are critical to the quality of classifiers founded through data mining methods. To extract useful information and diagnose the tumor, a hybrid of K-means and support vector machine (K-SVM) algorithms is developed. The K-means algorithm is utilized to recognize the hidden patterns of the benign and malignant tumors separately. The membership of each tumor to these patterns is calculated and treated as a new feature in the training model. Then, a support vector machine (SVM) is used to obtain the new classifier to differentiate the incoming tumors. Based on 10-fold cross validation, the proposed methodology improves the accuracy to 97.38%, when tested on the Wisconsin Diagnostic Breast Cancer (WDBC) data set from the University of California – Irvine machine learning repository. Six abstract tumor features are extracted from the 32 original features for the training phase. The results not only illustrate the capability of the proposed approach on breast cancer diagnosis, but also shows time savings during the training phase. Physicians can also benefit from the mined abstract tumor features by better understanding the properties of different types of tumors.

1. Introduction

Cancer is a major health problem in the United States. While conclusive data is not yet available, it was estimated that the number of new cancer cases in 2012 would approach 1,639,910 while the number of cancer deaths would reach at 577,190 (Siegel, Naishadham, & Jemal, 2012). Among the estimated new cancer cases in 2012, breast cancer was the most commonly diagnosed cancer among women, accounting for 29% of estimated new female cancer cases (790,740 cases) (Siegel et al., 2012). Diagnosing the tumors has become one of the trending issues in the medical field.

With the development of information technology, new software and hardware provide us ever growing ways to obtain mass descriptive tumor feature data and information on cancer research. Traditionally, breast cancer was predicted based on the mammography by radiologists and physicians. In 1994, ten radiologists were asked to analyze and interpret 150 mammograms to predict the tumor types in the breasts (Elmore, Wells, Lee, Howard, & Feinstein, 1994). Although the value of using mammograms was proven, the variability of the radiologists' interpretations caused a low accuracy of prediction. From their study, 90% of radiologists recognized fewer than 3% of cancers. Today, more and more technologies are utilized for collecting and analyzing the data. It is difficult for physicians to learn every detailed cancer feature from the large volume of cancer cases. Therefore, data analysis methodologies have become useful assistants for physicians when making cancer diagnosis decisions.

To increase the accuracy and handle the dramatically increasing tumor feature data and information, a number of researchers have turned to data mining technologies and machine learning approaches for predicting breast cancer. Data mining is a broad combination tool for discovering knowledge behind large scale data, and it has been shown to be highly applicable in the real world. In 1995, data mining and machine learning approaches were embedded into a computer-aided system for diagnosing breast cancer (Wolberg, Street, & Mangasarian, 1995); and a fuzzy-genetic approach to breast cancer diagnosis was proposed by Pena-Reyes and Sipper (1999). The results of their research showed that data mining technologies were successfully implemented in cancer prediction, and the traditional breast cancer diagnosis was transferred into a classification problem in the data mining domain. The existing tumor feature data sets were classified into malignant and benign sets separately. By figuring out a classifier to separate the two types of tumors, a new incoming

* Corresponding author. Tel.: +1 607 777 5935; fax: +1 607 777 4094.
E-mail address: yoons@binghamton.edu (S.W. Yoon).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.044

tumor could be predicted, based on the historical tumor data, by evaluating the classifier.

In the literature, data mining techniques were applied for diagnosing cancer based on tumor feature data. As the number of descriptive tumor features increases, the computational time increases rapidly as well. In this research, to deal with the large number of tumor features, methodologies for recognizing tumor patterns and extracting the necessary information for breast cancer diagnosis are studied. The objective of this paper is to find an efficient and accurate methodology to diagnose the incoming tumor type using data mining techniques. As the tumor features can be described as much detail as possible, the redundant information leads to a larger computation time for tedious calculation but without significant contribution to the final classifier. In this case, the basic requirement of cancer diagnosis is not only the accuracy but also the time complexity. With consideration of time efficiency, how to mine and extract the necessary information from the tremendous data sets, filter the features and predict the classification of the new tumor cases with high accuracy has become a new issue. Previously, Nezafat, Tabesh, Akhavan, Lucas, and Zia (1998) proposed sequential forward search and sequential backward search to select the most effective combination of features for obtaining a multilayer perceptron neural network to classify tumors. F-score (Chen & Lin, 2006) for determining the DNA virus discrimination was introduced for selecting the optimal subset of DNA viruses for breast cancer diagnosis using support vector machines (SVM) (Huang, Liao, & Chen, 2008). Akay (2009) proposed a SVM-based method combined with feature selection for breast cancer diagnosis. By using F-score (Chen & Lin, 2006) for measuring the feature discrimination, a time consuming grid search for the best parameter setting combination on diagnosis accuracy was conducted to select the optimal subset of the original tumor features for training by SVM (Akay, 2009). Prasad, Biswas, and Jain (2010) tried different combinations of heuristics and SVM to figure out the best feature subset for SVM training instead of the exhaustive search. Their results not only showed an improvement on cancer diagnosis accuracy, but reduced the computation time for the training significantly because of the deduction on feature space dimension. A defect was noted that those methods used the training accuracy as a criterion to evaluate different feature combinations. In other words, exhaustive training on different feature subsets was used to obtain the optimal subset with the best diagnosis accuracy, which was not time efficient. Thus, K-means algorithm as an unsupervised learning algorithm is proposed to extract tumor features in this paper to avoid the iterative training on different subsets. Since K-means algorithm clusters the original feature space by unsupervised learning, all of the individual feature information can be preserved in a more compact way for the following one-time training instead of multiple training pilots on different feature subsets. A membership function is developed to get the compact result of K-means algorithm ready for training by support vector machine (SVM), which was shown the high accuracy on breast cancer diagnosis (Bennett & Blue, 1998). Therefore, K-means algorithm and SVM are proposed to be hybrid for breast cancer diagnosis in this research.

The remainder of this paper is organized as follows: In Section 2, the SVM and feature selection and extraction methods are reviewed. K-means algorithm are brought to discuss for pattern recognition. The new approach based on feature extraction is proposed in Section 3. The experimental results are summarized in Section 4. In Section 5, the conclusion of this research is presented.

2. Literature review

Data mining (DM) is one of the steps of knowledge discovery for extracting implicit patterns from vast, incomplete and noisy data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996); it is a field with the confluences of various disciplines that has brought statistical analysis, machine learning (ML) techniques, artificial intelligence (AI) and database management systems (DBMS) together to address issues (Venkatadri & Lokanatha, 2011). Fig. 1 shows the importance of data mining in the knowledge discovery framework and how data is transferred into knowledge as the discovery process continues. Classification and clustering problems have been two main issues in the data mining tasks. Classification is the task of finding the common properties among a set of objects in a database and classifying them into different classes (Chen, Han, & Yu, 1996). Classification problems are closely related to clustering problems, since both put similar objects into the same category. In classification problems, the label of each class is a discrete and known category, while the label is an unknown category in clustering problems (Xu & Wunsch, 2005). Clustering problems were thought of as unsupervised classification problems (Jain, Murty, & Flynn, 1999). Since there are no existing class labels, the clustering process summarizes data patterns from the data set. Usually breast cancer has been treated as a classification problem, which is to search for a optimal classifier to classify benign and malignant tumors.

In the previous research, SVM was one of the most popular and widely implemented data mining algorithm in the domain of cancer diagnosis. The first part of the literature review introduces the basic concept and some implementations in the related areas of cancer diagnosis and predicting cancer survivability. Next, two main data pre-processing steps of data mining (feature extraction and selection) are reviewed to eliminate the redundant features and provide an efficient feature pattern for the classification model. Last but not the least important, the implementation of canonical K-means algorithm in this domain is summarized.

2.1. Support vector machine

Support vector machine is a class of machine learning algorithms that can perform pattern recognition and regression based on the theory of statistical learning and the principle of structural risk minimization (Idicula-Thomas, Kulkarni, Kulkarni, Jayaraman, & Balaji, 2006). Vladimir Vapnik invented SVM for searching a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin (Cortes & Vapnik, 1995). The margin was defined by the distance of the hyperplane to the nearest of the positive and negative examples (Platt,

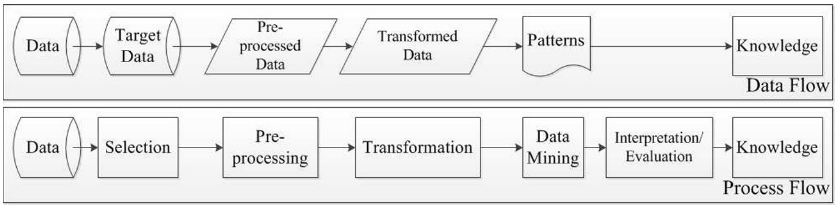


Fig. 1. An overview of knowledge, discovery, and data mining process (Fayyad et al., 1996).

1998). SVM has been widely used in the diagnosis of diseases because of the high accuracy of prediction. SVM generated a more accurate result (97.2%) than decision tree based on the Breast Cancer Wisconsin (Original) Dataset (Bennett & Blue, 1998). In the research for diagnosing breast cancer developed by Akay (2009), SVM provided 98.53%, 99.02%, and 99.51% for 50–50% of training-test partition, 70–30% of training-test partition, and 80–20% of training-test partition respectively based on the same previous data set which contained five features after feature selection by a genetic algorithm. In this research, the features were selected based on the rank of feature discrimination and the testing accuracy on different combinations of feature subsets using grid search and SVM, which requires high computational time and resources. In other words, to get the optimal parameter settings and feature subsets, the SVM trained the input iteratively until the optimal accuracy was obtained. The feature selection algorithm not only reduced the dimension of features but also eliminated the noisy information for prediction. Polat and Güneş (Polat & Güneş, 2007) proposed least square support vector machine (LS-SVM) for breast cancer diagnosis based on the same data set with accuracy of 98.53%. The main difference between LS-SVM and SVM was that LS-SVM used a set of linear equations for training instead of solving the quadratic optimization problem. By improving the training process, the calculation became simpler; however, feature selection was not combined in this research. Another SVM with a linear kernel was applied for the classification of cancer tissue (Furey et al., 2000), based on different data sets with more than 2,000 types of features.

2.2. Feature extraction and selection

Even when the same data mining approach is applied to the same data set, the results may be different since different researchers use different feature extraction and selection methods. It is important that the data is pre-processed before data mining is applied so that redundant information can be eliminated or the unstructured data can be quantified by data transformation. Theoretical guidelines for choosing appropriate patterns and features vary for different problems and different methodologies. Indeed, the data collection and pattern generation processes are often not directly controllable. Therefore, utilizing feature extraction and selection is the key to simplifying the training part of the data mining process and improving the performance without changing the main body of data mining algorithms (Platt, 1998).

Feature extraction, also called data transformation, is the process of transforming the feature data into a quantitative data structure for training convenience. The common features can be subdivided into the following types (Chidananda Gowda & Diday, 1991):

(1) Quantitative features:
  (a) continuous values (e.g., weight);
  (b) discrete values (e.g., the number of features);
  (c) interval values (e.g., the duration of an activity).
(2) Qualitative features:
  (a) nominal (e.g., color);
  (b) ordinal.

A generalized representation of patterns, called symbolic objects, was defined by a logical conjunction of events (Chidananda Gowda & Diday, 1991; Jain et al., 1999). These events link values and features to represent the abstract pattern of the data. Based on current research, feature extraction still focuses on transferring the data into a quantified data type instead of recognizing new patterns to represent the data.

The value of feature selection was illustrated by Jain and Zongker (1997). According to their research, feature selection played an important role in large scale data with high dimensional feature space, while it also had some potential pitfalls for small scale and sparse data in a high dimensional space. With high dimensional input, feature selection could be used for eliminating unnecessary information for training to reduce the overall training time while maintaining the original accuracy. Feature selection is mainly based on the performance of different feature combinations which means that to obtain the best combination, each possible combination of features would need to be evaluated. Among the different approaches for feature selection, genetic algorithm (GA) is one of the most popular. With the high dimension of feature space, GA provides a relatively good methodology of selecting features in a short time compared with other algorithms. It may not guarantee the optimal like the branch and bound method, but it performs well from the results and the time consumption perspective. GA was introduced into the feature selection domain by Siedlecki and Sklansky (1989). In the GA approach, a given feature subset is represented as a binary string (“chromosome”) of length n, with a zero or one in position i denoting the absence or presence of feature i in the set. Note that n is the total number of available features. A population of chromosomes is maintained. Each chromosome is evaluated to determine its “fitness,” which determines how likely the chromosome is to survive and breed into the next generation. New chromosomes are created from old chromosomes by the processes of (1) crossover where parts of two different parent chromosomes are mixed to create offspring, and (2) mutation where the bits of a single parent are randomly perturbed to create a child (Prasad et al., 2010).

Based on SVM classifier, three approaches, including GA, ant colony optimization (ACO) and particle swarm optimization (PSO), were utilized for selecting the most important features in the data set to be trained by the classification model (Prasad et al., 2010). The PSO-SVM showed the best results with 100% accuracy while GA-SVM provided 98.95% accuracy based on the Wisconsin Diagnostic Breast Cancer (WDBC) data set. In their research, the evaluation function of the optimization heuristic contained the training and validation accuracy, which means the whole data set had to be trained and validated to obtain the accuracy to evolve the algorithm to the next generation. The time it took to complete this part was not recorded by the researchers. Mu and Nandi developed a new SCH-based classifier for detecting the tumors and a hybridized genetic algorithm to search for the best feature combination to get a better result (Mu & Nandi, 2008). Feature selection has become another issue which cannot be ignored to get a more accurate result. Either getting rid of redundant information or reconstructing new patterns to represent the data is the objective of feature extraction and selection. This issue will also be solved by the proposed methodology in this paper.

2.3. K-means algorithm

Traditionally, K-means algorithm was for unsupervised learning clustering problems. It was not often utilized for predicting and classification problems but it was a good method for recognizing a hidden pattern from the data set (Jain et al., 1999). In the existing literature, the K-means algorithm is seldom used for diagnosing diseases directly which have been treated as a classification problem, but is implemented for exploring hidden patterns in the data. Constrained K-means algorithms were utilized for pattern recognition by Bradley, Bennett, and Demiriz (2000). The algorithm added K constraints in order to avoid empty clusters or a cluster with very few members to reduce the number of clusters. The algorithm was tested with the breast cancer diagnosis data set for clustering the benign and malignant tumors, which proposed another way of thinking about using clustering algorithms for recognizing patterns of data. Since their research was not focused on how to figure out a practical approach for determining the number of clusters on the data set, the number of clusters for the test on the breast cancer diagnosis data set was set to two arbitrarily for malignant and benign tumors. To discover the hidden patterns of breast cancer, determining the number of patterns should not be ignored. Also, the bridge between supervised and unsupervised learning for improving breast cancer diagnosis needs to be explored more.

To reduce the training set dimension, some researchers have started to combine clustering algorithms and classifier models in machine learning areas. Dhillon, Mallela, and Kumar (2003) used a hybrid clustering algorithm to group similar text words for achieving faster and more accurate training on the task of text classification. A variant of K-means algorithm, Fuzzy C-Means clustering algorithm was introduced to select training samples for SVM classifier training (Wang, Zhang, Yang, & Bu, 2012). Through the Fuzzy C-Means clustering algorithm, similar training samples were clustered and a subset of the training samples in the same cluster was selected for SVM classifier training.

From the previous discussion, several approaches have been utilized for breast cancer diagnosis based on classification and clustering. Yet in recent years, the amount of available data (both features and records) has increased dramatically. Traditional methodologies show their disadvantages on large scale data set. Although using meta-heuristics for feature selection reduces the number of features, the exhaustive enumeration on different feature subsets costs high computation time for different pilot training. The clustering algorithm, especially K-means algorithm, does not require a exhaustive search on feature selection, instead, it provides a good deduction on number of training samples without any contribution for feature selection and extraction. In this study, a hybrid of K-means algorithm and SVM is to condense the existing feature space to reduce the computational cost for SVM training and maintain a similar diagnosis accuracy.

3. Methodology

3.1. Data description

To implement the method for this research, a data set of Wisconsin Diagnostic Breast Cancer (WDBC) from the University of California – Irvine repository has been used. The WDBC data set was donated by Wolberg, in 1995. The data set contains 32 features in 10 categories for each cell nucleus, which are radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2/area – 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" – 1). For each category, three indicators are measured: mean value, standard error, and maximum value as shown in Table 1. Those different measurements are treated as different features in the data set. Since different features are measured in different scales, the error function will be dominated by the variables in large scale. Thus, to remove the effect of different scales, normalization is required before training. Totally, 569 instances have been collected with the diagnosed cancer results.

Table 1
Summary of data attributes.

Attributes	Measurement (Range)		
	Mean	Standard error	Maximum
Radius	6.98–28.11	0.112–2.873	7.93–36.04
Texture	9.71–39.28	0.36–4.89	12.02–49.54
Perimeter	43.79–188.50	0.76–21.98	50.41–251.20
Area	143.50–2501.00	6.80–542.20	185.20–4254.00
Smoothness	0.053–0.163	0.002–0.031	0.071–0.223
Compactness	0.019–0.345	0.002–0.135	0.027–1.058
Concavity	0.000–0.427	0.000–0.396	0.000–1.252
Concave points	0.000–0.201	0.000–0.053	0.000–0.291
Symmetry	0.106–0.304	0.008–0.079	0.157–0.664
Fractal dimension	0.050–0.097	0.001–0.030	0.055–0.208

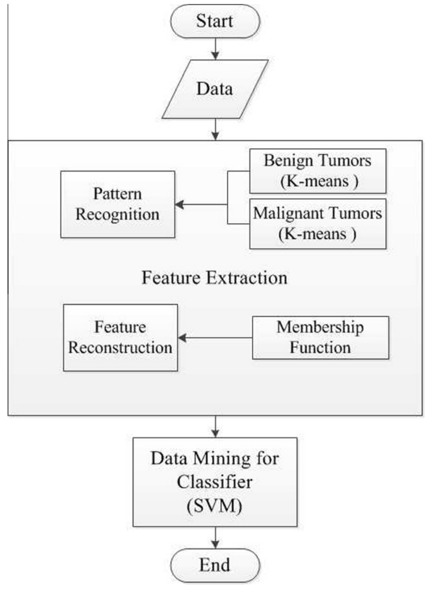


Fig. 2. General data mining framework.

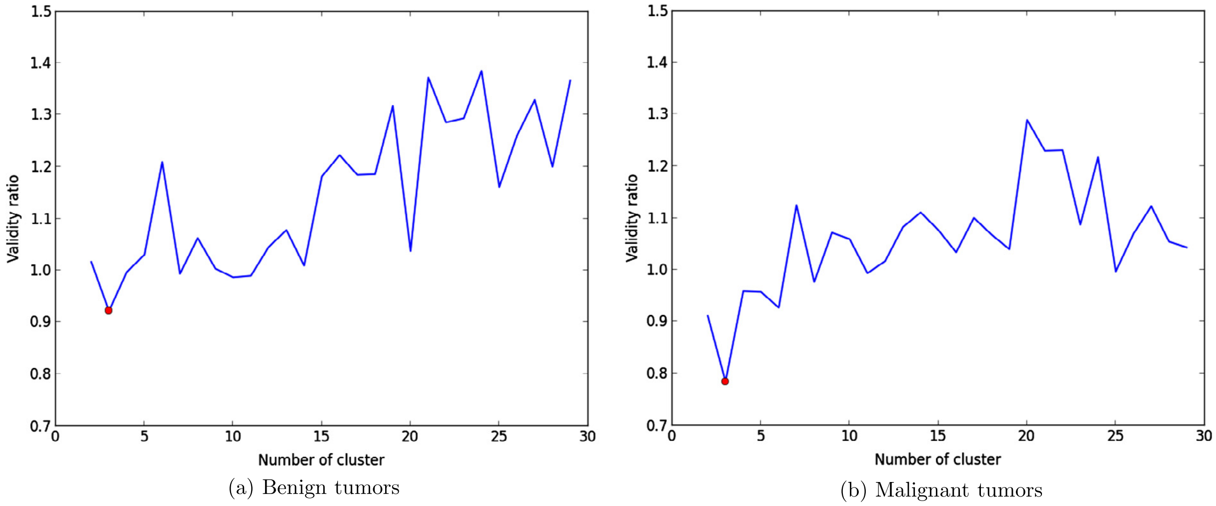
3.2. Feature extraction and selection

Fig. 2 shows the general steps of the proposed methodology. The objective of the breast cancer problem is to predict the property of a new tumor (malignant or benign). The proposed method hybridizes K-means algorithm and SVM (K-SVM) for breast cancer diagnosis. To reduce the high dimensionality of feature space, it extracts abstract malignant and benign tumor patterns separately before the original data is trained to obtain the classifier.

To recognize the patterns, feature extraction is employed. Inheriting the idea of symbolic objects, the K-means algorithm is used for clustering tumors based on similar malignant and benign tumor features respectively. A K-means problem can be formulated by using an optimization problem to minimize the overall distance between cluster centroids and cluster members as follows (Jain, 2010):

$$\min_{\mu_1,\mu_2,\ldots,\mu_K} \sum_{k=1}^{K}\sum_{i\in S_k}\|X^i-\mu_k\|^2 \qquad (1)$$

where k denotes the cluster index, S_k denotes the kth cluster set, μ_k is the centroid point in cluster S_k, which is also treated as the symbolic tumor of the cluster, and K is the total number of the clusters. It is important to normalize the data point for eliminating the effect of the different feature scales. To train the centroids used to construct the cluster, the K-means algorithm repeatedly adapts the centroid location for reducing the euclidean distance. There are two approaches for determining the number of clusters: (1) the physicians' experience and (2) a similarity measurement. The



(a) Benign tumors

(b) Malignant tumors

Fig. 3. Determine K for tumors.

second approach is applied to cluster similar tumor patterns. The similarity metrics to evaluate the clustering quality are shown in Eqs. (2) and (3) as follows:

$$d_{avg} = \frac{\sum_{k=1}^{K}\sum_{i\in S_k}\sqrt{\sum_{j=1}^{F}\left(X_j^i - X_j^{\mu_k}\right)^2}}{N} \tag{2}$$

$$d_{min} = \min\left[\sqrt{\sum_{j=1}^{F}\left(X_j^{\mu_{k_1}} - X_j^{\mu_{k_2}}\right)^2}\right], \quad \forall k_1 \neq k_2 \tag{3}$$

where d_{avg} is the average distance of each member i to the centroid μ_k in same cluster S_k, d_{min} represents the minimum distance between any two centroids, X_j^i denotes the jth input element of member i, $X_j^{\mu_k}$ denotes the jth input element of centroid μ_k, N is the total number of data points, and F is the dimension of an input vector. The optimal number of clusters (K^*) is obtained by getting the minimum validity ratio (θ) using Eq. (4) as follows:

$$K^* = \arg\min_{K}\theta = \arg\min_{K}\frac{d_{avg}}{d_{\min}} \tag{4}$$

where θ is the validity ratio for evaluating different numbers of clusters. This function is used to find an acceptable number of clusters to recognize the potential hidden patterns of benign and malignant tumors respectively. By searching for the minimum validity ratio, the average distance of each member to its cluster centroid (d_{avg}) is decreased, while the minimum distance between any two cluster centroids (d_{min}) is increased. In other words, the identified cluster is pushed itself to be compact and forced to be isolated from others. When K is close to the number of data points, it does not show several patterns with a large number of members, in other words, the density of clusters obtained by K-means algorithm is low, which is more similar to the original data. Thus, the smaller range of $K(2 \leqslant K \leqslant 30)$ is tried to find the local minimum number of clusters to represent the compact patterns of both types of tumors. The cluster center is considered as a new symbolic tumor of that cluster. In this case, the scale of the original data set has been reduced by the symbolic objects labeled by malignant and benign. The different trials for benign and malignant tumors are shown in Fig. 3. From the curve, three is chosen for the number of clusters for benign and malignant tumor sets separately since it is the local minimum (red point in the figure)[1] in the range of K from two to 30. Each cluster represents a specific tumor pattern. Each cluster centroid symbolizes the symbolic tumor of that cluster. To show the patterns clearly, the patterns for benign and malignant tumors are projected on three dimensions, which are shown in Fig. 4. The different color codes index different clusters, and the purple rectangle is the symbolic tumor of the cluster.

After recognizing the malignant and benign tumor patterns, several symbolic tumors have been formed in both the malignant and benign data sets. The similarity between the untested tumor and the symbolic tumors plays an important role for diagnoses. Therefore, a membership function is developed for measuring the similarity of the original data point and the symbolic tumors to show the fuzzy membership of the point to the identified patterns. The membership function is as follows:

$$f_c\left(X_j^i\right) = \begin{cases} 1 - \frac{|X_j^{\mu_c} - X_j^i|}{max\left|X_j^{\mu_c} - X_j^n\right|} & \text{if } \left(min\left(X_j^n\right) \leqslant X_j^i \leqslant max\left(X_j^n\right), \ \forall n \in S_c \right. \\ 0, & \text{otherwise;} \end{cases} \tag{5}$$

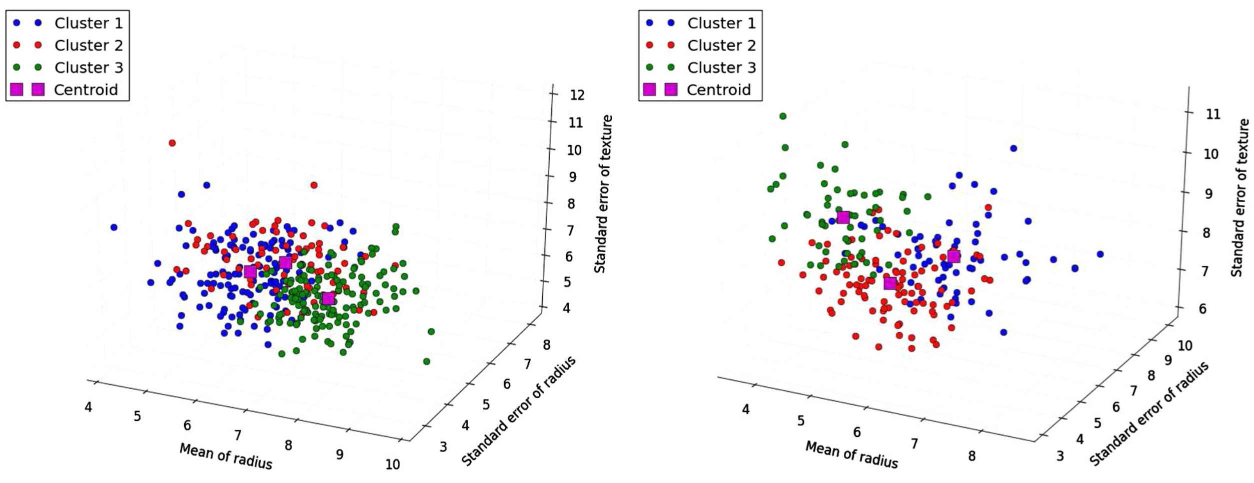
$$p_c = \frac{1}{F}\sum_{j=1}^{F} f_c\left(X_j^i\right), 1 \leqslant c \leqslant K^m + K^b \tag{6}$$

where c denotes the index of new pattern, X_j^i denotes the jth feature of the original input i, $X_j^{\mu_c}$ denotes the jth feature of centroid μ_c for cluster S_c obtained by previous K-means algorithm, and K^m and K^b are the numbers of malignant and benign patterns found by K-means algorithm respectively. With this membership function, the similarity between a tumor and the detected patterns is measured to show how well the tumor is fitted for the detected patterns. Therefore, in this approach the new pattern obtained by K-means is treated as the new abstract features of tumors. The new feature is different from the previous one that contains only one feature; it is a profile tumor pattern, which is condensed information combining different previous features. Thus, the feature space dimension is reduced. The value of the new feature represents the similarity between the tumor and the pattern. The boundary between benign and malignant tumors is determined by training the SVM based on these new features.

3.3. Data mining for classifier

Since the dimension of the feature space has been reduced, and the data set with new features has been rebuilt, traditional

[1] For interpretation of color in Fig. 3, the reader is referred to the web version of this article.



(a) $K = 3$ benign clusters

(b) $K = 3$ malignant clusters

Fig. 4. Tumor clusters.

machine learning algorithms can be applied here. SVM is used for obtaining precise classification because of its advantage of accuracy. The generalized SVM model is revised for this problem to search the classifier as follows (Cortes & Vapnik, 1995; Akay, 2009):

$$\text{maximize}_{\alpha} \quad \left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right] \tag{7}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_j = 0, \quad 0 \leqslant \forall \alpha_i \leqslant L. \tag{8}$$

where x is the training vector, y is the label associated with the training vectors, α is the parameters vector of classifier hyperplane, K is a kernel function for measuring the distance between the training vector x_i and x_j, and L is a penalty parameter to control the number of misclassification. For example, if L is infinity, the classifier gives an infinite penalty on misclassification to forbid misclassification from happening. A higher L gives a higher accuracy on training data; at the same time it consumes more time to obtain the classifier. A lower L provides more flexibility on the classifier on the tolerance of error. In this case, since the dimension of new features has been reduced to six, the different kernel function does not significantly affect the results. In this case, sigmoid kernel function has been applied in the SVM algorithm.

4. Experimental results

As stated in the previous discussion, the K-SVM is tested on the WDBC data set using 10-fold cross validation. The diagnosis accuracy is maintained at 97.38%, which is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative. To compare the performance with traditional SVM, the area under curves (AUC) of the K-SVM and SVM based on 10-fold cross-validation are shown in Fig. 5, which shows the stable good performance of the K-SVM and K-SVM keeps the high accuracy as SVM does. Even though the K-SVM reduces dimensionality of input feature space from 32 to six, the high prediction accuracy is maintained.

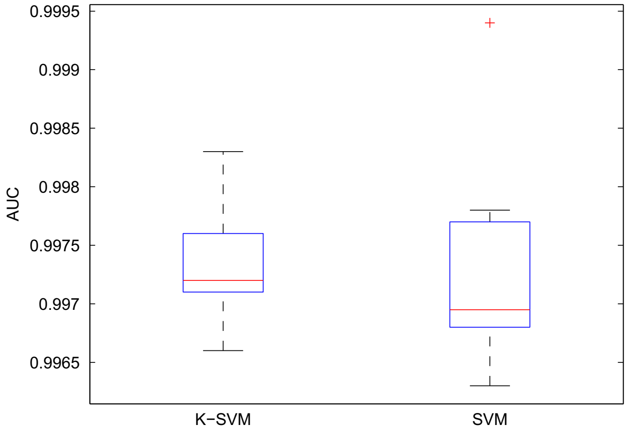


Fig. 5. AUC comparison.

In terms of accuracy, the proposed method provides stable and high prediction quality. Compared to the previous experimental results by Prasad et al. (2010), the K-SVM not only maintains similar accuracy but also refines and reduces the number of features as shown in Table 2 based on the 10-fold cross validation method and WDBC data set for the convenience of comparison with result from Prasad et al. (2010). The computational time for feature selection was not mentioned by Prasad et al. (2010). However, since Prasad's proposed method needs to calculate accuracy during training and validation to evaluate different possible feature combinations, it should consume more time. The advantage of the K-SVM is that it does not require feature selection during the training and validation phases, but, rather, the K-SVM reduces the input scale by transforming the original data into a new format. From the computation time perspective, the proposed method reduces the training time significantly by decreasing the number of the

Table 2
Result comparison.

	Feature space dimension	Accuracy (%)
K-SVM	6	97.38
ACO-SVM (Prasad et al., 2010)	15	95.96
GA-SVM (Prasad et al., 2010)	18	97.19
PSO-SVM (Prasad et al., 2010)	17	97.37

Table 3
CPU time for classification.

	Feature space dimension	CPU time (seconds)
K-SVM	6	0.0039
SVM	30	15.8913

input features. The computation time is compared with the traditional SVM algorithm in Table 3, which shows the importance of selecting and extracting the features.

Based on the final results, the proposed approach to recognize the patterns of benign and malignant tumors and reconstruct the new input feature with the membership of new patterns based on original data provides better accuracy than other approaches. The tumor pattern recognition actually suggests a new thought on how to define features for the tumors. Feature selection filters the properties that are not main factors for the diagnosis of cancer. Therefore, feature extraction and selection play an important role in the cancer diagnoses. The K-SVM reduces the input feature space dimension and reconstructs the format of the features for supporting the machine learning algorithm to optimize the classifier. With the effort of feature extraction and selection, the training time to obtain the classifier has been reduced, and the classifier accuracy is improved since the noisy information has been eliminated.

5. Conclusion and future work

In this paper, the K-SVM based on the recognized feature patterns has been proposed. It can be competitively compared with traditional data mining methods in cancer diagnosis. For the phase of feature extraction, the traditional methods of extracting useful information are not used. Instead, clustering is used to extract the symbolic tumor objects to represent tumor clusters. The similarity between the incoming tumor and the symbolic tumor is measured as the membership of the pattern to predict if the new case can be diagnosed as cancer or not. K-means clustering algorithm is utilized to recognize and obtain the patterns of malignant and benign tumors respectively. These patterns are reconstructed as the new abstract tumor features for the training phase. The pre-processing steps (feature extracting and selection) provide a highly effective and compact feature set for the machine learning algorithms to train the classifier. According to the result, the K-SVM reduces the computation time significantly without losing diagnosis accuracy. However, by using K-means algorithm, the training sample size has not been decreased by filtering the similar samples in the data, which could be a potential way of reducing the dimension of training set further.

The K-SVM in this paper still follows the knowledge discovery framework shown in Fig. 1, and more and more knowledge of cancer would be discovered by data mining methodologies. Moreover, features and descriptive data would continue to be collected in the future. The feature extraction and selection would still be a challenge for the researchers. Meanwhile, the large data set with missing values is another challenge to be conquered in terms of computation time. In this research, the data set used is a complete data set without missing values. However, the implementation of this method in a large scale sparse data set would be a direction in which to expand in the future. So far in the cancer diagnosis domain, there is no general rule for selecting the number of patterns for two types of tumors. In the future, a way of determining the number of symbolic tumors should be developed. This would reduce the time for feature extraction and allow physicians to draw inspiration for understanding the tumors based on the patterns gained by data mining approaches.

Acknowledgement

This study was supported by the Watson Institute for Systems Excellence at State University of New York at Binghamton. The authors also wish to thank anonymous reviewers and colleagues who have given valuable comments to improve this paper.

References

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications, 36*, 3240–3247.

Bennett, K. P., & Blue, J. A. (1998). A support vector machine approach to decision trees. In *Proceedings of IEEE world congress on computational intelligence* (pp. 2396–2401). Anchorage, AK: IEE.

Bradley, P. S., Bennett, K. P., & Demiriz, A. (2000). Constrained K-means clustering. *Technical report microsoft research redmond*. WA, USA.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*, 866–883.

Chen, Y.-W., & Lin, C.-J. (2006). Combining svms with various feature selection strategies. In *Feature extraction* (pp. 315–324). Berlin Heidelberg: Springer.

Chidananda Gowda, K., & Diday, E. (1991). Symbolic clustering using a new dissimilarity measure. *Pattern Recognition, 24*, 567–578.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Dhillon, I. S., Mallela, S., & Kumar, R. (2003). A divisive information theoretic feature clustering algorithm for text classification. *The Journal of Machine Learning Research, 3*, 1265–1287.

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., & Feinstein, A. R. (1994). Variability in radiologists interpretations of mammograms. *New England Journal of Medicine, 331*, 1493–1499.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine, 17*, 37–54.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics, 16*, 906–914.

Huang, C.-L., Liao, H.-C., & Chen, M.-C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Systems with Applications, 34*, 578–587.

Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K., & Balaji, P. V. (2006). A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in escherichia coli. *Bioinformatics, 22*, 278–284.

Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*, 153–158.

Jain, A. K. (2010). Data clustering: 50 Years beyond k-means. *Pattern Recognition Letters, 31*, 651–666.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR), 31*, 264–323.

Mu, T., & Nandi, A. K. (2008). Breast cancer diagnosis from fine-needle aspiration using supervised compact hyperspheres and establishment of confidence of malignancy. The 16th European Signal Processing Conference, EUSIPCO, Lausanne, Switzerland.

Nezafat, R., Tabesh, A., Akhavan, S., Lucas, C., & Zia, M. (1998). Feature selection and classification for diagnosing breast cancer. In *Proceedings of international association of science and technology for development international conference* (pp. 310–313).

Pena-Reyes, C. A., & Sipper, M. (1999). A fuzzy-genetic approach to breast cancer diagnosis. *Artificial Intelligence in Medicine, 17*, 131–155.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. In *Advances In Kernel Methods – Support Vector Learning* (pp. 185–208). Cambridge, MA, USA: MIT Press.

Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing, 17*, 694–701.

Prasad, Y., Biswas, K., & Jain, C. (2010). Svm classifier based feature selection using ga, aco and pso for sirna design. In *Proceedings of the first international conference on advances in swarm intelligence* (pp. 307–314).

Siedlecki, W., & Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters, 10*, 335–347.

Siegel, R., Naishadham, D., & Jemal, A. (2012). Cancer statistics, 2012. *CA: A Cancer Journal for Clinicians, 62*, 10–29.

Venkatadri, M., & Lokanatha, C. R. (2011). A review on data mining from past to the future. *International Journal of Computer Applications, 15*, 19–22.

Wang, X.-Y., Zhang, X.-J., Yang, H.-Y., & Bu, J. (2012). A pixel-based color image segmentation using support vector machine and fuzzy c-means. *Neural Networks, 33*, 148–159.

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1995). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative Cytology and Histology, 17*, 77–87.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*, 645–678.