A Survey on Cancer Prediction Using Data Mining Techniques

Angeline Prasanna G[1], P.C Sakthipriya[2]

[1]Head of the Department, Department of Computer Application & Information Technology, Kaamathenu Arts & Science College, Sathyamangalam, Tamilnadu, India.
[2]Research Scholar, Department of Computer Application & Information Technology, Kaamathenu Arts & Science College, Sathyamangalam, Tamilnadu, India

---

Abstract: Conventional semi-supervised clustering approaches have several shortcomings, such as (1) not fully utilizing all useful must-link and cannot-link constraints, (2) not considering how to deal with high dimensional data with noise, and (3) not fully addressing the need to use an adaptive process to further improve the performance of the algorithm. In this paper, we first propose the transitive closure based constraint propagation approach, which makes use of the transitive closure operator and the affinity propagation to address the first limitation. Then, the random subspace based semi-supervised clustering ensemble framework with a set of proposed confidence factors is designed to address the second limitation and provide more stable, robust and accurate results.

Next, the adaptive semi-supervised clustering ensemble framework is proposed to address the third limitation, which adopts a newly designed adaptive process to search for the optimal subspace set. Finally, we adopt a set of nonparametric tests to compare different semi-supervised clustering ensemble approaches over multiple datasets. The experimental results on 20 real high dimensional cancer datasets with noisy genes and 10 datasets from UCI datasets and KEEL datasets show that (1) The proposed approaches work well on most of the real-world datasets. (2) It outperforms other state-of-the-art approaches on 12 out of 20 cancer datasets, and 8 out of 10 UCI machine learning datasets.

Index Terms: Clustering ensemble, semi-supervised clustering, clustering,

1. INTRODUCTION:

Cancer is the most dreadful disease which is a collection of related diseases. In every cancer types, some of the body's cells begin to divide involuntarily and spread in to surrounding tissue.

Likewise, breast cancer arises from breast cell tissues. Breast, like any part of the body, consists of billions of microscopic cells. These cells start to multiply uncontrollably causing breast cancer. There are mainly 2 types of breast cancer, a ductal carcinoma is the most common cancer where cancer begins in the milk duct and the cancer that begins in the lobules are called lobular carcinoma[1]. At the ancient times, the diagnosis and treatment proves to be fatal without efficient techniques. At each stage, the vigorosity of this disease and the death rate is increasing. So there is a need of early detection of breast cancer to reduce the death rate. So automated computer aided detection is inaveitable.

The detailed detection of the images is done using feature extraction and texture extraction techniques. This itself opens another research area as a wide variety of techniques have been used for segmentation, feature extraction, enhancement. Done mainly by wavelet techniques, clustering, using GLCM matrix etc which are described clearly in the related works. So the ultimate aim of this survey is to provide different enhancement, detection and classification techniques for early breast cancer detection.

Semi-supervised clustering [1]-[3] is an important sub-field of clustering and is widely applied in different areas, such as image processing [4]-[5], multimedia [6], pattern recognition [7][8] and bioinformatics [9][10]. For example, Biswas et al. [4] applied the constrained clustering algorithm for image analysis. Liu et al. [5] proposed a novel semi-supervised matrix decomposition approach, and applied it to image processing and document clustering.

In this paper, we focus on constrained clustering, which belongs to the class of semi-supervised clustering approaches. Constrained clustering integrates a set of must-link constraints and cannot-link constraints into the clustering process. The must-link constraint means that two data samples should belong to the same cluster, while the cannot-link constraint means that two data samples cannot be assigned to the same cluster. Traditional constrained clustering approaches have two limitations: (1) They do not consider how to make full use of must-link constraints and cannot-link constraints; (2) Some methods do not take into account how to deal with high dimensional data with noise.

In order to address the limitations of traditional constrained clustering approaches, we first propose a transitive closure based constraint propagation approach, which not only expands the constraint set using transitive closure, but also adopts

the label propagation approach to disseminate the pairwise constraints. Then, we design a Random subspace based SEMIsupervised Clustering Ensemble framework (RSEMICE) to integrate the clustering solutions, which are obtained through different transitive closure based constraint propagation processes from multiple datasets, into a unified solution. Next, we propose an adaptive semi-supervised clustering ensemble framework, which adopts a newly designed adaptive process to optimize the subspace set and improve the performance of RSEMICE further. Finally, a set of nonparametric tests are used to compare different semi-supervised clustering ensemble approaches over multiple datasets.

The contribution of the paper is threefold. First, we propose a random subspace based semi-supervised clustering ensemble framework to provide more stable, robust and accurate results based on a set of confidence factors. Second, an adaptive process is designed to search for the optimal subspace set and improve the performance of RSEMICE. Third, we adopt a set of nonparametric tests to compare constrained clustering approaches over multiple datasets.

2 RELATED WORKS:

Semi-supervised clustering is one of the important research directions in the area of data mining, which is able to make use of prior knowledge, such as pairwise constraints or a small amount of labeled data, to guide the search process and improve the quality of clustering.

A number of semi-supervised clustering approaches have been previously proposed, which can be divided into five categories. The approaches belonging to the first category focus on designing new kinds of semi-supervised clustering algorithms, such as semi-supervised hierarchical clustering, semisupervised kernel mean shift clustering , semi-supervised maximum margin clustering , semi-supervised linear discriminant clustering , semi-supervised subspace clustering, semi-supervised matrix decomposition , semisupervised information-maximization clustering , active semi-supervised fuzzy clustering , semi-supervised kernel fuzzy c-means , semi-supervised clustering framework based on discriminative random fields , semi-supervised fuzzy clustering based on competitive agglomeration , semi-supervised clustering corresponding to spherical Kmeans and feature projection , constrained clustering based on Minkowski weighted K-means , semi-supervised nonnegative matrix factorization based on constraint propagation [24], semi-supervised kernel mean shift clustering [56], semisupervised clustering based on seeding [65], and so on.

3 EXPERIMENTS:

The performance of the proposed approaches are evaluated using cancer gene expression profiles in Table 1 (where n denotes the number of data samples, m denotes the number of attributes, and k denotes the number of classes), which are challenging datasets with high dimensionality and small sample size. For example, the Garber-2001 dataset contains 66 samples, and each sample has 4553 dimensions. Conventional semi-supervised clustering approaches cannot obtain satisfactory results on thisdataset. For another example, the Ramaswamy-2001 dataset has only 190 samples with very high dimension, which are assigned to 14 classes.

In this case, traditional constraint clustering methods cannot be effectively applied to this dataset due to the large number of classes. In summary, these datasets can be used to more thoroughly explore the performance bounds of the semi-supervised clustering ensemble approaches.

TABLE 1

Dataset	Index	Source	n	m	k
Alizadeh-2000-v3(o)	S1	[49]	62	4096	4
Armstrong-2002-v2	S2	[49]	72	2194	3
Bredel-2005	S3	[49]	50	1739	3
Chen-2002	S4	[49]	179	85	2
Chowdary-2006	S5	[49]	104	182	2
Dyrskjot-2003	S6	[49]	40	1203	3
Garber-2001	S7	[49]	66	4553	4
Golub-1999-v2	S8	[49]	72	1877	3
Khan-2001	S9	[49]	83	1069	4
Lapointe-2004-v1	S10	[49]	69	1625	3
Lapointe-2004-v2	S11	[49]	110	2496	4

A summary of the real-world cancer datasets (where n denotes the number of data samples, m denotes the number of attributes and k denotes the number of classes).

We also investigate the performance of our proposed approaches against a number of state-of-the-art semi-supervised clustering ensemble algorithms on 10 UCI datasets summarized in Table 2. As we can see, they present great difficulties for the task of clustering. For example, the movement libras dataset contains 360 data samples with 15 classes. Due to the challenging properties (small number of samples and large number of classes) of the dataset, the issue of how to efficiently utilize the information of pairwise constraints to enhance the performance and stabilityof our clustering approach becomes acute.

The effect of the transitive closure operator:

In order to explore the effect of the transitive closure operator, RSEMICE with transitive closure (RSEMICE) is compared with the random subspace based semi-supervised clustering ensemble framework without considering transitive closure (RSEMICE-TC) on all the cancer datasets in Table 1 based on ARI. Table 4 shows the results obtained by RSEMICE and RSEMICE-TC. It can be seen that RSEMICE obtains better results on 19 out of 20 datasets. For example, RSEMICE achieves the best performance with an ARI value of 0.7737 on the Garber-2001 dataset (S7), which is 0.2765 greater than that obtained by RSEMICE-TC. The possible reason is that transitive closure is able to enlarge the pairwise constraint set and increase the number of pairwise constraints. This additional information will in turn be useful to improve the performance. As a result, the transitive closure operator plays an important role in RSEMICE, and if it is missing, the effectiveness of our algorithm will be reduced for most of the datasets. In addition, transitive closure is a general technique which can be incorporated into different pairwise constraintbased approaches to improve their performance.

4. Classification techniques:

Support Vector Machine Support Vector Machine is one type of learning system algorithm, which is used to perform classification more accurately. SVM used for two class classifier. The essence of SVM is hyper plane also known as “decision boundary or decision surface”. This hyper plane separates the positive and negative of training data sample. ¬**Advantages•** SVM is easy to extended, useful to pattern reorganization, formulated quadratic optimization problem. ¬ **Disadvantages•** It is suitable for real valued space. • For binary classification, it allow only two classes and for multiple class classification, it apply several strategies • Hyper plane is very hard to use by the user .

K- Nearest Neighbor Learning KNN is another one important learning method to classify the simple data set. Comparing with other learning classifier approaches k-nn method is quite effective. Here the learning occurs when test examples need classified. 3.6.1 Algorithm[k,d,D] • compute the distance between d and every example in D; • choose the k examples in D that are nearest to d, denote the set by P(D); • assign d the class that is the most frequent class in P[6].

5. DETECTION AND CLASSIFICATION METHODS Detection and classification is done with a wide variety of techniques. Some of the techniques are below

5.1 Detection Techniques:

5.1.1 Wavelet Neural Network The wavelets has got both spatial and domain characteristics and thus can be used to find out the abnormalities. The ANN is the best tool for the diagnosing the pathological images especially of cancers and precancers. By combining the wavelet theory and neural network we will get best detection method for breast cancer.

5.1.2 Curvelet transforms It is a new image representation. The main feature is that it has efficient image coding, geometrical features, optimal object representation with edges, image reconstruction, optimal sparse representation of wave propagators than wavelets. Here feature vector is taken as curvelet transform coefficients.

5.1.3 Contourlet transform The minute geometrical features of the images such as smooth contours are detected using contourlet transform. This has got multiscale and directional properties. Through this a mammographic image is decomposed in to various directional subband at different scales using 2D filter bank.

5.1.4 GLDM For an image GLDM calculates the GLDM probability density functions. Commonly used for extracting statistical textural features of a mammographic image. Five texture features are defined: Contrast, Angular Second moment, Entropy, Mean and Inverse Difference Moment from each density functions.

5.1.5 Gabor Filter and Histogram Equalisation Image quality is increased using histogram equalization method. Gabor filter used to remove the noisy signals present in the image. A 2-D Gabor function is a Gaussian modulated by a sinusoid. It is a non-orthogonal wavelet.

fAnalysis of detection and classification techniques of mammographic images Here the different techniques for detection and classification are discussed and compared. J. Dheebaet.al[8] proposed a Particle Swarm Optimised Wavelet Neural Network (PSOWNN) to classify and detect breast cancer. WNN posses both wavelet and neural network properties, here the ROI detection is done through Global thresholding technique. WNN along with PSO proves to be best for classification as it decreased false negatives and false positives. From the mammographic images the laws texture energy measures are extracted with an abnormality detection algorithm. The textural features are extracted through a windowing operation and by convolution kernels applied to ROI.216 mammographic images obtained from mammographic screening centres have been used. Advantage is that this method increased convergence of back propagation algorithm error and the disturbances in learning are avoided. K. Subashini et. al[11] proposed breast cancer detection through ultrasound images. Wavelet domain techniques i.e, DWT translates the images in to wavelet coefficients to remove noise. Here the noise representing coefficients are suppressed and image features are enhanced. Segmentation is done with active contour model and the texture features are extracted using auto covariance coefficients. The back propagation neural network are used for classification which found better in performance than linear classifiers.

References:

[1] X. Zhu, “Semi-supervised learning literature survey”, Department of Computer Sciences, University of Wisconsin-Madison, 2008.

[2] X. Zhu, A.B. Goldberg, “Introduction to semi-supervised learning”, Morgan & Claypool, 2009.

[3] Every Women Counts, Resource for Health Professionals. ☒

[4] National Breast Cancer accounts. ☒

[5] A Review On Breast Abnormality Segmentation And Classification Techniques☒

[6] S. Shanthi, and V. MuraliBhaskaran ,’Computer Aided System for Detection and Classification of Breast Cancer’, International Journal of Information Technology, Control and Automation (IJITCA) Vol.2, No.4, October 2012

[7] Neeta Jog, Arvind Pandey, ’Implementation of Segmentation and Classification Techniques for Mammogram Images’, IOSR Journal of Engineering (IOSRJEN), Vol. 05, Issue 02 (February. 2015)

[8] S. Deepa, Dr.V.SubbiahBharathi, ‘Textural Feature Extraction and Classification of Mammogram Images using CCCM and PNN’, IOSR Journal of Computer Engineering .

[9]Thiyagarajan C, Dr.K.Anandhakumar, Dr A.Bharathi “Diabetes Mellitus Diagnosis based on Transductive Extreme Learning Machine”, International Journal of Computer Science and Information Security (IJCSIS),ISSN 1947-5500, Volume 15, Number 6, June 2017.