Semantic Risk Analysis Model Cancer Data Prediction

S.B. Jayabharathi, S. Manjula, E. Tamil Selvan, P. Vengatesh and
V. Senthil Kumar

Abstract--- In this paper breast cancer risk assessment model can assess whether a user is at a high risk of developing breast cancer disease or not and confirm a breast cancer high-risk group. Because the etiology of breast cancer disease is different in different country and region, the existing risk assessment model is only adaptive to certain countries and regions. And the parameters of these models are fixed, so these models have poor generality. Aiming at these problems, the paper puts forward a new breast cancer risk assessment model named as Shrink. Using the idea of social network, Shrink constructs a medical social network to show the similarity among user, and uses group division algorithm to divide the network into breast cancer high-risk group and low-risk group. The parameters of this model can be set according to the needs of the breast census, and these parameters can be directly acquired through questionnaire, therefore Shrink has good generality. Moreover, under the uncertain classification standard, Shrink adopts a new classification method to discover breast cancer high-risk group.

In addition Shrink model solves the poor generality of the existing model. The factors used by the SVM model are acquired through questionnaire, and the model doesn't depend on certain factors. Every country can set factors according to the need of the breast census. Under uncertain classification standards, Shrink uses group division algorithm to discover breast cancer high-risk group. Also, through experiment analysis, the SVM model itself has good judgment function, and the model is better than KNN model. Therefore, improve results the prevention and control of breast cancer.

Keywords--- Breast Cancer Detection, Group Division Algorithm, SVM Model, Social Network, KNN Classification Model.

I. INTRODUCTION

Hussain Fatakdawala et al [1] describe a the presence of lymphocytic infiltration (LI) has been correlated with nodal metastasis and tumor recurrence in HER2+ breast cancer (BC), making it important to study LI. The ability to detect and quantify extent of LI could serve as an image based prognostic tool for HER2+ BC patients. Lymphocyte segmentation in H & E-stained BC histopathology images is, however, complicated due to the similarity in appearance between lymphocyte nuclei and cancer nuclei.

Additional challenges include biological variability, histological artifacts, and high prevalence of overlapping objects. Although active contours are widely employed in segmentation, they are limited in their ability to segment overlapping objects. In this paper, proposed a segmentation scheme (EMaGACOR) that integrates Expectation Maximization (EM) based segmentation with a geodesic active contour (GAC). Additionally, a novel heuristic edge-path algorithm exploits the size of lymphocytes to split contours that enclose overlapping objects. In addition, the scheme differs from supervised classifier detection methods

S.B. Jayabharathi, UG Scholar, Department of Computer Science and Engineering, K.S.R College of Engineering, Tiruchengode, Namakkal. E-mail:bjayabharathi2127@gmail.com

S. Manjula, UG Scholar, Department of Computer Science and Engineering, K.S.R College of Engineering, Tiruchengode, Namakkal. E-mail:sweetmanju110@gmail.com

E. Tamil Selvan, UG Scholar, Department of Computer Science and Engineering, K.S.R College of Engineering, Tiruchengode, Namakkal. E-mail:elagotamil110@gmail.com

P. Vengatesh, UG Scholar, Department of Computer Science and Engineering, K.S.R College of Engineering, Tiruchengode, Namakkal. E-mail:vengatesh.pcse@ksrce.ac.in

V. Senthil Kumar, Assistant Professor, Department of Computer Science and Engineering, K.S.R College of Engineering, Tiruchengode, Namakkal. E-mail:senthilkmr777@gmail.com,mkkumz@gmail.com

that are encumbered by the need for a large number of annotated training samples.

Hui Kong, Metin Gurcan [2] proposed an integrated framework consisting of a novel supervised cell-image segmentation algorithm and a new touching-cell splitting method. The segmentation algorithm learns a most discriminant color space in a linear discriminant framework so that the extracted local Fourier transform features can achieve optimal classification (segmentation) in it.

In addition, proposed an efficient LFT feature extraction scheme to speed up the segmentation process. In the touching-cell clump splitting step, proposed an novel strategy in that the touching-cell clump is differentiated from the non-touching cells beforehand and only the touching-cell clump is split by a new iterative splitting algorithm.

Ricardo Gutiérrez et al [3] describe A reliable determination of clinically meaningful Regions of Interest (RoIs) in medical images is at the very base of strategies for selective image analysis, adaptive delivering of image data and clever compression algorithms. A proper determination of these RoIs would allow to concentrate any processing effort on specific image areas, relevant within a particular context. This fundamental statement would improve the processing performance in applications such as medical education, medical training, decision support systems, virtual microscopy and telepathology. One of the most challenging issues in histopathological images regarded the fact that semantic interest is related to similarity, no matter whether these regions are neighbors or not. This drawback was herein dealt with a graph-based image segmentation algorithm, which in contrast to previous approaches, was capable of capturing perceptually important regions such as tissue distribution. As illustrated in Figure 6 regions obtained with the proposed strategy are perceptually more consistent and coherent with what the expert set.

Payel Ghosh, Sameer Antani et al [4] presents a review of online systems for content-based medical image retrieval (CBIR). The objective of this review is to evaluate the capabilities and gaps in these systems and to determine ways of improving relevance of multi-modal (text and image) information retrieval in the iMedline system, being developed at the National Library of Medicine (NLM). Seven medical information retrieval systems: Figure search, BioText, GoldMiner, Yale Image Finder, Yottalook, Image Retrieval for Medical Applications (IRMA), and iMedline have been evaluated here using the system of gaps. Not all of these systems take advantage of the visual information contained in biomedical literature as figures and illustrations.

László SzilágyiSándor M et al [5] describe Automated brain MR image segmentation is a challenging pattern recognition problem that received significant attention lately. The most popular solutions involve fuzzy c-means (FCM) or similar clustering mechanisms. Several improvements have been made to the standard FCM algorithm, in order to reduce its sensitivity to Gaussian, impulse, and intensity non-uniformity noises. The paper presents a modified FCM-based method that targets accurate and fast segmentation in case of mixed noises. The method extracts a scalar feature value from the neighborhood of each pixel, using a context dependent filtering technique that deals with both spatial and gray level distances. These features are clustered afterwards by the histogram-based approach of the enhanced FCM algorithm. Results were evaluated based on synthetic phantoms and real MR images. Test experiments revealed that the proposed method provides better results compared to other reported FCM-based techniques. The achieved segmentation and the obtained fuzzy membership values represent excellent support for deformable contour model based cortical surface reconstruction methods.

C. Iswarya, R. et al [6] describe a novel method for unsupervised change detection in multi-temporal satellite images using Gaussian mixture model (GMM) with spatial information is proposed. This approach is based on three steps. Firstly, the difference image between two Synthetic

Aperture Radar (SAR) images of the same area taken at two different times is obtained using the standard log–ratio operator.

Susmita Ghosh, et al [7] propose an unsupervised context-sensitive technique for change-detection in multi temporal remote sensing images. Here a modified self-organizing feature map neural network is used. Each spatial position of the input image corresponds to a neuron in the output layer and the number of neurons in the input layer is equal to the number of features of the input patterns. The network is updated depending on some threshold value and when the network converges, status of output neurons depict a change-detection map. To select a suitable threshold of the network, a correlation based and an energy based criteria are suggested. The proposed change-detection technique is unsupervised and distribution free. Experimental results, carried out on two multispectral and multitemporal remote sensing images, confirm the effectiveness of the proposed approach. In remote sensing applications, change-detection is the process of identifying differences in the state of an object or phenomenon by analyzing a pair of images acquired on the same geographical area at different times

II. BREAST CANCER RISK ASSESSMENT MODEL

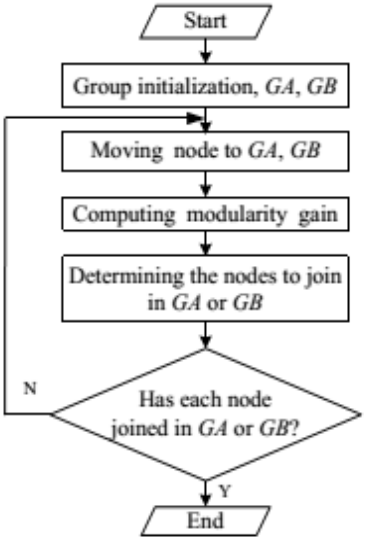
Medical research discovers that the people who live in a similar environment have similar habits and customs, the behavior way, etc., so the probability of suffering the same kind of disease is also high. Therefore, the similarity among people can be as a reference to divide the people into different groups. The idea of dividing group according to the similarity let us think of community structure in social network.

Medical Social Network Construction

The network is usually represented as graphs. We use G= (V, E) to represent network with a set of nodes V and edges E V V*. E denotes the weight of edges. V denotes people. In our model, the weight represents the similarity of rrfi. Some people who have similar rrfi may have the same kind of health conditions. The similarity value represents the closing degree of rrfi between two people. Some approaches, such as Euclidean distance and cosines similarity measure, have been proposed to calculate the similarity of document data for recommendation system.

Group Division

By utilizing the module optimization algorithm of community detection, we design the group division algorithm. In medical social network, there are breast cancer patients and hale user. The hale people mean that these people haven't been examined through clinical diagnosis, but these patients are confirmed breast cancer cases.



Shrink Realization Algorithm

In Shrink realization algorithm, when different rrf is used, the network construction and division method are the same. The algorithm will be iterated to run for RRF which includes orrfs. In the beginning of the algorithm, the numbers of hale people, the numbers of breast cancer patients, and the numbers of related risk factors need to be initialized. In the iteration process, the inputs parameters initializations are different for each execution.

Input: N numbers of hale people, P numbers of breast cancer patients, and rrfm ę RRF (RRF has o rrfs)
Output: GA, GB
1: For x=1: (N+P)
2: For y=x+1: (N+P)
3: Compute (,)
s x yrrfm
4: Endfor
5: Endfor
6: Network G is built with
srrfm
7: Group initialization, according to rules, selects
8: two nodes pi and Nj ,pięGA, NjęGB,
9: For s=1 to (N+P)
10: If (s! = pi and s! = Nj)
11: Move s to GA, compute QGA , Move s to
12: GB, compute QGB ,
13: if
QGA > QGB
14: s joins in GA
15: Else
16: s joins in GB
17: Endif
18: Endif
19: Endfor

III. SUPPORT VECTOR MACHINE AND KNN CLASSIFICATION MODEL

SVM Model

SVM have attracted a great deal of attention in the last decade and actively tested to various domains applications. SVMs are mostly used for learning classification, regression or ranking function. SVM are based on statistical learning theory and structural risk minimization principal and have the intent of determining the location of decision boundaries also known as hyperplane that produce the optimal separation of classes.

SVM is the most robust and exact classification technique, there are many problems. The data analysis in SVM is based on convex quadratic programming, and it is computationally costly, as solving quadratic programming methods require large matrix operations as well as time consuming numerical computations.

Pseudo-code for the SVM (T) is stated below:

- Step1: ComputeClassFrequency(T);
- Step 2: if OneClass or FewCases
    - Return a leaf; Create a decision node N;
- Step 3: ForEach Attribute A ComputeGain(A);
- Step 4: N.test = AttributeWithBestGain;
- Step 5: if N.test is continuous Find Threshold;
- Step 6: For Each T' in the splitting of T
- Step 7: if T' is Empty Clild of N is a leaf else Child of N = DecisionTree(T');
- Step 8: ComputeErrors of N; Return N.

K-Nearest Neighbor (KNN)–Techniques

KNN is a supervised learning algorithm which classifies new data based on minimum distance from the new data to the K nearest neighbor. The proposed work has used Euclidean Distance to define the closeness.

Pseudo-code for the KNN classifier is stated below:

- Step 1: Input: D={(x1,c1),….(xn,cn)} x=(x1,….., xn) new instance to be classified
- Step 2: For each labeled instance (xi,ci) Calculated (xi, x)
- Step 3: Orderd(xi, x) from lowest to highest, (i=1, …,N)
- Step 4: Select the K nearest instances to x : Dx K
- Step 5: Assign to x the most frequent class in Dx K

K-Nearest Neighbors Algorithm

The k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest

training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

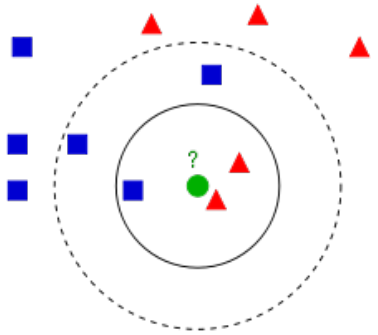


Fig. 1: k-NN Classification

DATA Reduction

Data reduction is one of the most important problems for work with huge data sets. Usually, only some of the data points are needed for accurate classification. Those data are called the prototypes and can be found as follows:

Select the class-outliers, that is, training data that are classified incorrectly by k-NN (for a given k)

1. Separate the rest of the data into two sets: (i) the prototypes that are used for the classification decisions and (ii) the absorbed points that can be correctly classified by k-NN using prototypes. The absorbed points can then be removed from the training set.

Selection of class-outliers

A training example surrounded by examples of other classes is called a class outlier. Causes of class outliers include:

- random error
- insufficient training examples of this class (an isolated example appears instead of a cluster)

Missing important features (the classes are separated in other dimensions which we do not know). Too many training examples of other classes (unbalanced classes) that create a "hostile" background for the given small class

Class outliers with k-NN produce noise. They can be detected and separated for future analysis. Given two natural numbers, k>r>0, a training example is called a (k,r)NN class-outlier if its k nearest neighbors include more than r examples of other classes

CNN for Data Reduction

Condensed nearest neighbor (CNN, the Hart algorithm) is an algorithm designed to reduce the data set for k-NN classification.[17] It selects the set of prototypes U from the training data, such that 1NN with U can classify the examples almost as accurately as 1NN does with the whole data set. Given a training set X, CNN works iteratively:

1. Scan all elements of X, looking for an element x whose nearest prototype from U has a different label than x.
2. Remove x from X and add it to U
3. Repeat the scan until no more prototypes are added to U.

Use U instead of X for classification. The examples that are not prototypes are called "absorbed" points. It is efficient to scan the training examples in order of decreasing border ratio.[18] The border ratio of a training example x is defined as

$$a(x) = ||x'-y|| / ||x-y||$$

where ||x-y|| is the distance to the closest example y having a different color than x, and ||x'-y|| is the distance from y to its closest example x' with the same label as x.

The border ratio is in the interval [0,1] because ||x'-y|| never exceeds ||x-y||. This ordering gives preference to the borders of the classes for inclusion in the set of prototypes U. A point of a different label than x is called external to x. The calculation of the border ratio is illustrated by the figure on the right. The data points are labeled by colors: the initial point is x and its label is red. External points are blue and green. The closest to x external point is y. The closest to y red point is x' . The border ratio a(x)=||x'-y||/||x-y|| is the attribute of the initial point x. Below is an illustration of CNN in a series of figures. There are three classes (red, green and blue).

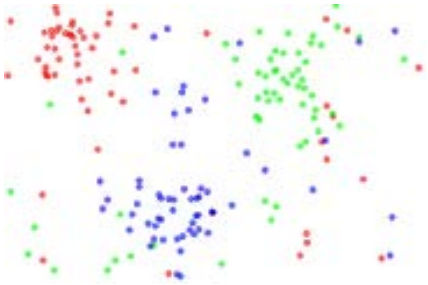


Fig. 2 shows initially there are 60 points in each class

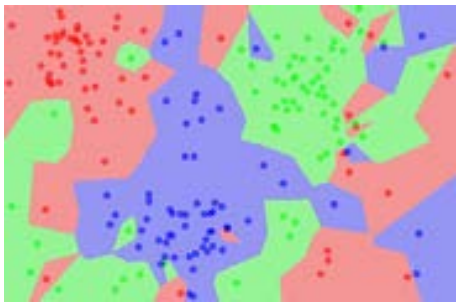


Fig. 3 shows the 1NN classification map: each pixel is classified by 1NN using all the data

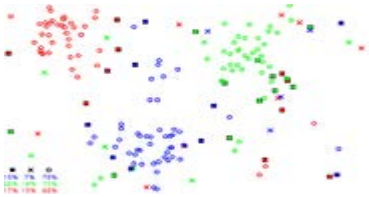


Fig. 4 shows the 5NN classification map. White areas correspond to the unclassified regions, where 5NN voting is tied (for example, if there are two green, two red and one blue points among 5 nearest neighbors)

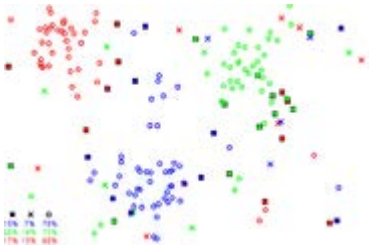
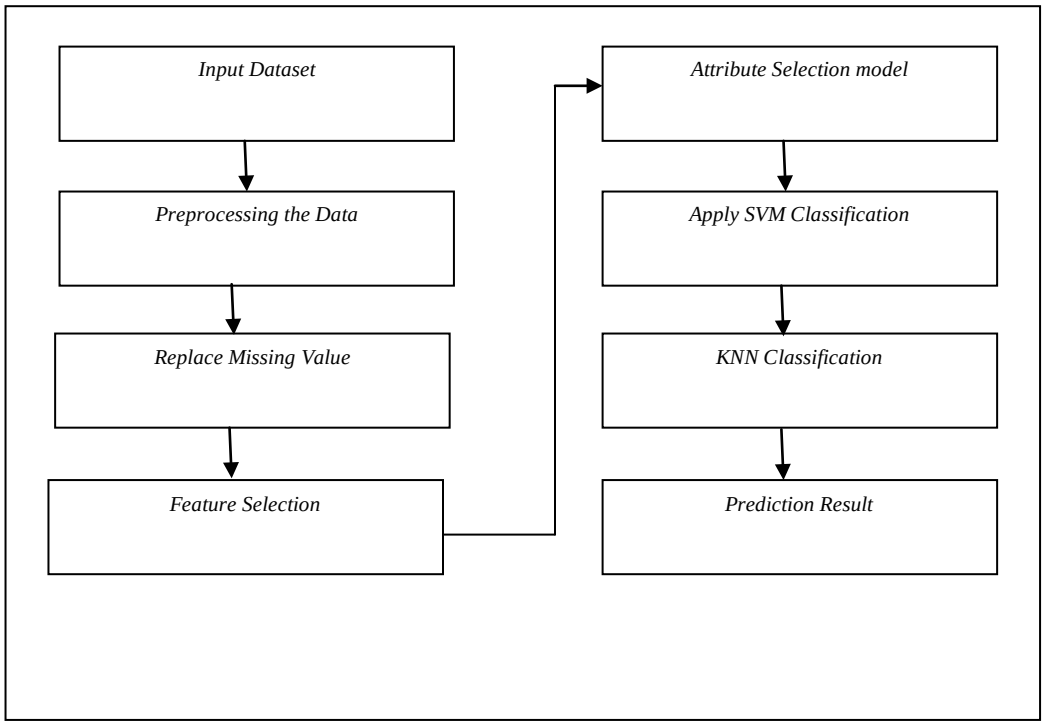


Fig. 5 shows the reduced data set. The crosses are the class-outliers selected by the (3, 2) NN rule (all the three nearest neighbors of these instances belong to other classes); the squares are the prototypes, and the empty circles are the absorbed points. The left bottom corner shows the numbers of the class-outliers, prototypes and absorbed points for all three classes. The number of prototypes varies from 15% to 20% for different classes in this example.

Fig. 6 shows that the 1NN classification map with the prototypes is very similar to that with the initial data set. The figures were produced using the Mirkes apple

In k-NN regression, the k-NN algorithm is used for estimating continuous variables. One such algorithm uses a weighted average of the k nearest neighbors, weighted by the inverse of their distance. This algorithm works as follows:

1. Compute the Euclidean or Mahalanobis distance from the query example to the labeled examples.
2. Order the labeled examples by increasing distance.
3. Find a heuristically optimal number k of nearest neighbors, based on RMSE. This is done using cross validation.
4. Calculate an inverse distance weighted average with the k-nearest multivariate neighbors.



IV. CONCLUSION

This paper deals with the results in the field of data classification obtained with Shrink Realization algorithm, SVM algorithm and KNN algorithm, and on the whole performance made known Shrink Realization algorithm when tested on cancer disease datasets, time taken to run the data for result is fast when compared to other algorithms. It shows the enhanced performance according to its attribute. Attributes are fully classified by this algorithm and it gives 80% of accurate result. In future, based on the experimental results the classification accuracy is found to be better using Navi basyes algorithm compare to other algorithms. From the above results Random Forest Tree algorithm plays a key role in shaping improved classification accuracy of a dataset.

REFERENCES

[1] P. Li, Y. Wang, Y. Tian, T.S. Zhou and J.S. Li, "An Automatic User-Adapted Physical Activity Classification Method Using Smartphones", IEEE Transactions on Biomedical Engineering, Vol. 64, No. 3, Pp.706-714, 2017.

[2] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology", Proc. 5th IEEE Int. Symp. Biomed. Imag., Pp. 284–287, 2008.

[3] H. Fatakdawala, J. Xu, A. Basavanhally, G. Bhanot, S. Ganesan, M. Feldman, J.E. Tomaszewski and A. Madabhushi, "Expectation–maximization-driven geodesic active contour with overlap resolution (emagacor): Application to lymphocyte segmentation on breast cancer histopathology", IEEE Transactions on Biomedical Engineering, Vol.57, No.7, Pp.1676-1689, 2010.

[4] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: An

integrated framework for supervised colortexture segmentation and cell splitting", IEEE Trans. Med. Imag., Vol. 30, No. 9, Pp. 1661–1677, 2011

[5] R. Gutirrez, F. Gmez, L. Roa-Pea, and E. Romero, "A supervised visual model for finding regions of interest in basal cell carcinoma images", Diagnostic. Pathol., Vol. 6, No. 26, 2011.

[6] C.R. Angel, D. Gloria, R. Eduardo and G. Fabio, "Automatic annotation of histopathological images using a latent topic model based on nonnegative matrix factorization", J. Pathol. Informat., Vol. 2, No. 2, 2011.

[7] P. Ghosh, S. Antani, L. Long and G. Thoma, "Review of medical image retrieval systems and future directions", Proc. 24th Int. Symp. Comput.-Based Med. Syst., Pp. 1–6, 2011.

[8] A. Kumar, J. Kim, W. Cai, M. Fulham and D. Feng, "Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data", J. Digital Imag., Vol. 26, No. 6, Pp. 1025–1039, 2013.

[9] X. Zhang, W. Liu, M. Dundar, S. Badve and S. Zhang, "Towards largescale histopathological image analysis: Hashing-based image retrieval", IEEE Trans. Med. Imag., Vol. 34, No. 2, Pp. 496–506, 2015

[10] J. Caicedo, F. Gonzalez and E. Romero, "A semantic content-based retrieval method for histopathology images", Information Retrieval Technology (Series Lecture Notes in Computer Science 4993), Berlin, Germany: Springer-Verlag, Pp. 51–60, 2008.