http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 6 (8): 1258-1263, 2007
ISSN 1812-5638

Feature Space Optimization in Breast Cancer Diagnosis Using Linear Vector Quantization

A. Punitha and T. Santhanam
P.G. and Research Department of Computer Science, D.G. Vaishnav College,
Arumbakkam, Chennai, India

Abstract: One of the major challenges in medical domain is the extraction of intelligible knowledge from medical diagnosis data. It is quite common among the researching community to apply Principal Component Analysis (PCA) for the extraction of prominent features and to use feature correlation method for redundant features removal. This paper discusses a three-phase approach selection technique to extract features for further usage in clinical practice for better understanding and prevention of superfluous medical events. In the first phase PCA is employed to extract the relevant features followed by the elimination of redundant features using the class correlation and feature correlation technique in phase two and in the final phase Learning Vector Quantization (LVQ) network is utilized for classification. The proposed method is validated upon Wisconsin Breast Cancer Database (WBCD), which is a very well known dataset obtained from the UCI machine-learning repository. The abridged feature set and classification accuracy are found to be satisfactory.

Key words: Correlation, learning vector quantization network, principal component analysis, relevance and redundancy

INTRODUCTION

Breast cancer is the most common cancer in women worldwide. Trends and Statistics indicate that one out of nine women will develop breast cancer in their lifetime and one out of 27 women die due to breast cancer (American Cancer Society, 2006).

Medical domain makes heavy use of databases to store information about patients, like patient's history, surveys and medical investigations from several devices. The medical databases contain data in a variety of formats like images, textual information, psychology reports, medical articles or various signals like Electro Cardio Gram (ECG), Electro Encephala Gram (EEG) etc. that are called features.

Currently, processing abilities fail to handle such high dimensional data due to numerical difficulties in processing, requirements in storage and transmission within a reasonable time. To reduce the computational time, it is a general practice to compress the data, reduce the noise and to extract the predominant features without sacrificing the efficiency.

Classification without feature selection is certainly not the best technique even for data sets with relatively a small number of features. This shows the so-called curse of dimensionality and the necessity for feature selection. Feature selection has been an active and fruitful field of research and development for several decades in the statistical pattern recognition, machine learning and data mining. It has been proven in theory and practice that reduction of dimension is effective in enhancing learning efficiency, increasing predictive accuracy and reducing model complexity.

Different feature selection methods can be broadly categorized into the wrapper model (Kohavi and John, 1997) and the filter model (Lei and Liu, 2004). The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected features. The filter model separates feature selection from classifier learning and selects feature subsets that are independent of any learning algorithm. It relies on various measures of the general characteristics of the training data such as distance, information, dependency and consistency (Dash *et al*., 2000). Search is another key problem in feature selection. Different search strategies such as complete, heuristic and random search have been studied to generate candidate feature subsets for evaluation (Dash and Liu, 1997). According to the availability of class labels, there are feature selection methods for supervised learning (Lei and Liu, 2004) and unsupervised learning (Kim *et al*., 2000).

In the case of classification problems, data are often given as a set of vectors with each element of each vector being a value of some feature (Krzysztof and Halina, 2006), $f_i \in F = \{f_1 \ldots f_k\}$. It can be assumed that if the

Corresponding Author: T. Santhanam, P.G. and Research Department of Computer Science, D.G. Vaishnav College, Arumbakkam, Chennai, India

features are real-valued, introducing a set of vectors $V = \{v_1 \ldots vn\} \subset R^k$, a set of classes C and the classifier $K:R^k \rightarrow C$. Obviously,

$$\forall vi \in V, j \in \{1,..., k\} \bullet v_{ij} \in f_j \quad (1)$$

Data reduction can be performed by selecting a subset $F' \subset F$. It is not easy to decide which features are to be considered so that the quality of classification made using the reduced space consisting of the features considered is efficient.

Both the filter and the wrapper approaches require a search procedure that iterates over the space of possible feature sets. Some basic strategies of feature selection are individual ranking (Kittler, 1978) forward search and backward search.

Univariate selection starts with an empty set $F_0 = \emptyset$. In each step, one best individually ranked feature f' is added,

$$F_n = F_{n-1} \cup \{f'\} \quad (2)$$

Where:

$$f' = \arg\max Q(f_i) \quad (3)$$
$$F_{n-1} \cap \{f_i\} = \emptyset$$

Where, $Q(f_i)$ is the quantitative criterion that measures the capability of the feature set F_i to discriminate between classes. But individual ranking does not take into account the existence of any dependencies between features and may therefore give poor results.

Forward search also starts with an empty set $F_0 = \emptyset$. In each step, one feature f' is added which maximizes the criterion Q together with previously selected features

$$F_n = F_{n-1} \cup \{f'\} \quad (4)$$

Where:

$$f' = \arg\max Q(F_{n-1} \cup \{f'\}) \quad (5)$$
$$F_{n-1} \cap \{f_i\} = \emptyset$$

Forward search takes into consideration at least some of the potential interdependencies between features, but the required feature set is constructed in a greedy manner, which may also produce sub optimal results (Pudil *et al*., 1994).

Backward s earch starts with the set of all features $F_0 = F$ and in each step it removes one feature f' which, when removed from the selected features set, maximizes the criterion Q,

$$F_n = F_{n-1} \setminus \{f'\} \quad (6)$$

Where:

$$f' = \arg\max Q(F_n = F_{n-1} \setminus \{f'\}) \quad (7)$$
$$F_{n-1} \cap \{f_i\} \neq \emptyset$$

Backward search is not only a greedy approach but also computationally complex than forward search, as it requires the criterion Q to be evaluated on a representation space of much higher dimensionality than the one used in forward search.

Recently, a pair wise selection strategy was proposed (Pekalska *et al*., 2005). Pair wise selection takes into consideration at least some possible interdependencies between features and has reasonable computational complexity. In this selection strategy, the selection procedure begins with an empty set $F_0 = \emptyset$. Then, in each step of the iterative process, the best pair of features is added to the set of selected features F_n,

$$F_n = F_{n-1} \cup \{f', f''\} \quad (8)$$

Where:

$$\{f', f''\} = \arg\max Q(F_n = F_{n-1} \cup \{f', f''\}) \quad (9)$$
$$i \neq j$$
$$F_{n-1} \cap \{f', f''\} = \emptyset$$

Existing feature selection methods mainly exploit two approaches: individual evaluation and subset evaluation (Blum and Langley, 1997). Methods of individual evaluation rank features according to their importance in differentiating instances of different classes and can only remove irrelevant features, as redundant features likely have similar rankings. Methods of subset evaluation search for a minimum subset of features that satisfies some goodness measure and can remove irrelevant features as well as redundant ones.

Pair-wise selection generally takes into account the possible relationship between the features and it is computationally more complex and redundant. So, in this study a new methodology is introduced for relevant feature selection and to eliminate redundancy as far as possible.

TRADITIONAL PCA

Feature selection is a dimensionality reduction technique that selects a subset of new features from the original set by means of some functional mapping keeping as much information in the data as possible.

Conventional PCA is a one of the frequently applied feature selection techniques that is based on extracting the axes on which the data shows the highest variability. Although this approach spreads out the data in the new basis and can be of great help in regression problems and unsupervised learning, there is no guarantee that the new axes are consistent with the discriminatory features in a classification problem. PCA applies second order methods that use the covariance structure in determining directions that restricts directions to those that are orthogonal.

PCA transforms the original set of features into a smaller subset of linear combinations that account for most of variance of the original set (Alexey *et al*., 2002).

The main idea of PCA is to determine the features, which explain as much of the total variation in the data as possible with as few of these features as possible. In PCA the authors are interested in finding a projection w:

$$Y = W^T X$$

Where:
Y = Transformed data point
W = Transformation matrix
X = Original data point

PCA can be done through eigenvalue decomposition of the covariance matrix S of the original data:

$$S=1/n\sum_{i=1}^{n} (X_i - m)\,(X_i - m)^T$$

Where:
n = The No. of instances
X_i = The i-th instance
m = The mean vector of the input data

The principal components can be computed using the following algorithm:

- Calculate the covariance matrix S from the input data.
- Compute the eigenvalues and eigenvectors of S and sort them in a descending order with respect to eigenvalues.
- Form the actual transition matrix by taking the predefined number of components (eigenvectors).
- Finally, multiply the original feature space with the obtained transition matrix, which yields a lower-dimensional representation.

The necessary cumulative percentage of variance explained by the principal axes should be consulted in order to set a threshold, which defines the number of components to be chosen. The threshold value has been carefully chosen keeping in mind the predominant features are not lost.

The salient features of PCA are: (1) it maximizes the variance of the extracted features (2) the extracted features are uncorrelated (3) it finds the best linear approximation in the mean-square sense and (4) it maximizes the information contained in the extracted features.

Though there are numerous advantages, one main problem with PCA is that, it gives more weightage to features with higher variabilities without taking into consideration whether they will be useful for further classification or not.

CORRELATION

The purpose of feature selection is to decide, which of the initial features are to be included in the final subset. If the number of features is n, then there will be 2^n possible subsets. The only way to discover the best subset of this is clearly to find the correlation between the variables (Lei and Liu, 2004).

To evaluate the goodness of features for classification correlation is adopted, as correlation analysis attempts to determine the degree of relationship between variables. The most well known measure is linear correlation coefficient δ for a pair of variables chosen as (X, Y).

$$\delta = \frac{N\sum XY - \sum(X)\sum(Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\times\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

If X and Y are completely correlated, δ takes the value of 1 or -1 and if X and Y are totally independent, δ is zero, as it is a symmetrical measure for two variables. On adopting the correlation between variables as a goodness measure for classification, a feature is said to be good if it is highly correlated with respect to the class.

Feature relevance: Based on the review of definitions for feature relevance (John *et al*., 1994), classified features into three disjoint categories, namely, strongly relevant, weakly relevant and irrelevant features (Kohavi and John, 1997). Let F be a full set of features, F_i a feature and $S_i = F - \{F_i\}$. These categories of relevance can be formalized as follows:

Definition 1 (Strong relevance) A feature F_i is strongly relevant if

$$P(C \mid F_i, S_i) \neq P(C \mid S_i)$$

Definition 2 (Weak relevance) A feature F_i is weakly relevant iff

$$P(C \mid F_i, S_i) = P(C \mid S_i) \text{ and}$$
$$\exists\, S'_i \subset S_i \text{ such that } P(C \mid F_i, S'_i) \neq P(C \mid S'_i)$$

Corollary 1 (Irrelevance) A feature F_i is irrelevant iff

$$\forall\, S'_i \subseteq S_i, P(C \mid F_i, S'_i) = P(C \mid S'_i)$$

Strong relevance of a feature indicates that the feature is always necessary for an optimal subset; it cannot be removed without affecting the original conditional class distribution. Weak relevance suggests that the feature is not always necessary but may become necessary for an optimal subset at certain conditions. Irrelevance indicates that the feature is not necessary at all. An optimal subset should include all strongly relevant features, none of irrelevant features and a subset of weakly relevant features. However, there is no hard and fast rule to decide which of the weakly relevant features to be considered. Therefore, it is essential to define feature redundancy among relevant features.

Feature redundancy: Notions of feature redundancy are normally in terms of feature correlation. It is widely accepted that two features are redundant to each other if their values are completely correlated. In reality, it may not be so straightforward to determine feature redundancy when a feature is correlated with a set of features.

At the heart of the correlation, it is a heuristic for evaluating the worth or merit of a subset of features. This heuristic takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. The hypothesis by (Mark and Lloyd, 1998) which the heuristic is stated as good feature subsets contain features highly correlated with (predictive of) the class yet uncorrelated with (not predictive of) each other.

In general, a feature is good if it is relevant to the class concept but is not redundant to any of the other relevant features. So on adopting the correlation between two variables as a goodness measure, the above definition becomes that a feature is good if it is highly correlated to the class but not to any other features.

MATERIALS AND METHODS

The proposed algorithm performs relevance and redundancy analysis by selecting only the predominant features. It employs PCA for relevance and correlation for redundancy and is shown in Fig. 1. On combining PCA with backward elimination, this achieves higher efficiency than pure sequential forward or backward selection.

On applying PCA, a feature F_i is said to be relevant if it is highly dominant with the principal axes. The features are selected only if PC $(F_i) > \gamma$ where γ is the relevance threshold (which can be determined by users) and then subjecting these selected relevant features for redundancy analysis.

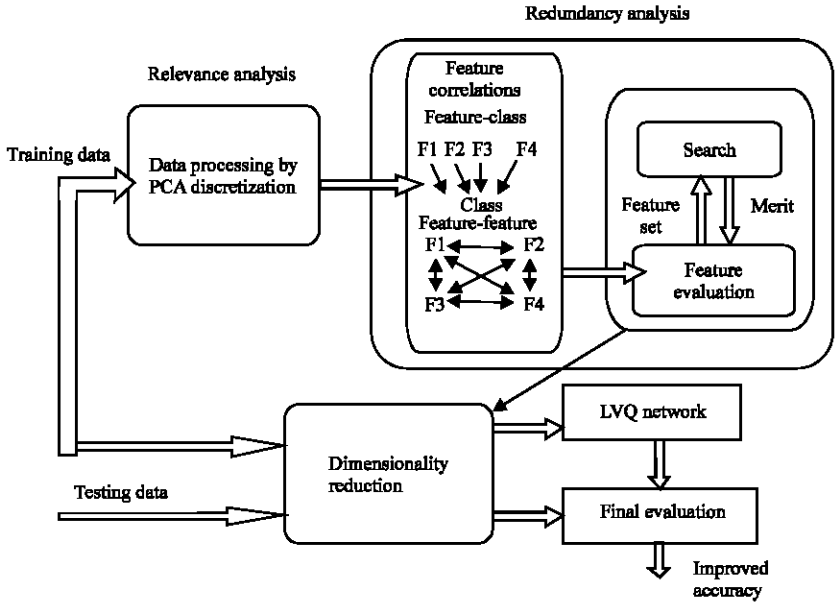

Fig. 1: Feature selection using PCA and correlation

From the obtained principal components, Feature Correlation (FC) between individual features is evaluated for redundancy analysis. When two features are completely correlated with each other any one of them can be eliminated, but it may be hard to determine which feature to be removed. Class Correlation (CC) is applied to both the features and the feature with a smaller CC value gets eliminated. Principal Component's Pair wise Correlation Based Filter (PCPCBF) is used in this study for not only determining the relevant features but also for removing the redundant ones. It involves two connected steps namely (i) Selecting a subset of relevant features by PCA and (ii) Selecting predominant features from relevant ones to produce a final subset. For a data set S with N features and C class, the algorithm finds a set of predominant features S_{best}.

ALGORITHM PCPCBF

$S = \{x_1 \ldots .x_n\}$ // Set of data
S_{best} // Best subset of features
γ = 80.7, δ = 75.4 // PCA and Correlation thresholds, respectively

(1) Calculate PCA(S);
(2) For I = 1 to N do begin
(3) If $PC(F_i) > \gamma$
Append F_i to S_l
(4) End
(5) Order S_l in decending order; according to $PC(F_i)$
(6) Fj = getFirstElement(S_l);
(7) Do begin
(8) Fi = getNextElement(S_l, Fj);
(9) if (F_j <> NULL)
(10) Do begin
(11) if (FC(i,j) >δ)
(12) if (CC(i) > CC(j))
(13) remove Fj from S_l;
(14) else
(15) remove Fi from S_l;
(16) F_j = getNextElement(S_l, F_j);
(17) End (Until (F_j == NULL);
(18) Fj = getNextElement(S_l, Fj);
(19) End until (F_j <> NULL);
(20) $S_{best} = S_l$;
(21) End

This procedure ensures that any feature that correlated with any other feature at the level of δ, will be evaluated in a pair-wise manner, while the features that have no significant correlation with any other feature will be evaluated individually. However this method is also sub optimal due to the approximations that are used for selecting features both in relevance and redundancy analysis.

RESULTS

Fine Needle Aspiration (FNA) of breast masses is the most non-invasive test that obtains information needed to evaluate malignancy. The Wisconsin Breast Cancer Database (WBCD) consists of 683 samples with 16 missing attributes (Blake and Merz, 2006) has been employed in this work. The database consists of nine features obtained from fine needle aspirates, each of which is ultimately represented as an integer value between 1 and 10. The measured variables are as follows: (1) Clump Thickness (X1) (2) Uniformity of Cell Size (X2) (3) Uniformity of Cell Shape (X3) (4) Marginal Adhesion (X4) (5) Single Epithelial Cell Size (X5) (6) Bare Nucleoli (X6) (7) Bland Chromatin (X7) (8) Normal Nucleoli (X8) and (9) Mitoses (X9) where 444 of the data set belong to benign and remaining 239 data is of malignant.

Vector Quantization is a standard statistical clustering technique, which seeks to divide the input space into areas that are assigned as codebook vectors. In Learning Vector Quantization network, target values are available for the input pattern and the learning is supervised. In training process, the output units are positioned to approximate the decision surfaces. After training, an LVQ net classifies an input vector by assigning it to the same class as the output unit that has its weight vector closest to the input vector. It takes the first 'm' training vectors and uses them as weight vectors; here m is taken as 2 and the remaining vectors are used for training, also the learning rate is taken as 0.1. Initialize the reference vectors randomly and assign the initial weights and class randomly. K-means clustering method can be adopted.

The efficiency is compared with general PCA and Class Correlation with all inputs and with PCPCBF algorithm. For PCPCBF, the PCA threshold is taken as 80.83 and 7 features are selected. Now on looking into pairwise inter correlation between PCA selected features for removing the redundant features, the 4 features selected are relevant but without redundancy.

To evaluate the proposed approach, a number of experiments were carried out with the WBCD, to find the efficiency and time taken to train the LVQ network with selected feature and with entire feature set. The efficiency obtained is 94.72 which classifies 647 out of 683 samples (Table 1). It is compared with the other techniques of reduction and classification (Kohavi and John, 1997).

Table 1: Comparison of feature selection accuracies by C4.5, ID3, RLF and PCPCBF

Parameters	Total No. of features	C4.5	RLF	ID3	PCPCBF
Features	10.00	7.0	5.70	9.1	4.00
Efficiency	94.57	95.4	95.14	96.5	94.72

From the Table 1, it is very clear that the proposed method has optimized the features to a greater level and without compromising on the efficiency.

CONCLUSIONS

Generally, there is degradation in the performance of many machine-learning algorithms when irrelevant features are present. Feature subset selection definition is to find a subset of features that maximizes the accuracy of a classifier by minimizing the time taken. Thus the proposed algorithm shows that features selected by considering relevance and redundancy not only increases the accuracy but also reduces the time taken to train the network. In future, the algorithm proposed in this work will be explored with high dimensionality databases.

REFERENCES

Alexey, T., P. Seppo, P. Mykola, B. Matthias and P. David, 2002. Eigenvector-based Feature Extraction for Classification. American Association for Artificial Intelligence (www.aaai.org). All Rights Reserved. Copyright ©.

American Cancer Society, 2006. Breast Cancer Facts and Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

Blake, C. and C. Merz, 2006. UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/~ mlearn/MLRepository.html.

Blum, A.L. and P. Langley, 1997. Selection of relevant features and examples in machine learning. Artif. Intell., 97: 245-271.

Dash, M. and H. Liu, 1997. Feature selection for classification. Intelligent data analysis: Annu. Int. J., 1 (3): 131-156.

Dash, M., H. Liu and H. Motoda, 2000. Consistency Based Feature selection. Proceedings of the 4th Pacific Asia Conference on Knowledge Discovery and Data Mining, Springer-Verlag, pp: 98-109.

John, G., R. Kohavi and K. Pfleger, 1994. Irrelevant feature and thesubset selection problem. In: Proceedings of the 11th International Conference on Machine Learning, pp: 121-129.

Kim, Y., W. Street and F. Menczer, 2000. Feature selection for unsupervised learning via volutionary search. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp: 365-369.

Kittler, J., 1978. Pattern Recognition and Signal Processing. The Netherlands: Sijhoff and Noordhoff, pp: 41-60.

Kohavi, R. and G.H. John, 1997. Wrappers for feature subset selection. Artif. Intell., 97: 273-324.

Krzysztof, M. and K. Halina, 2006. Correlation-based feature selection strategy in classification problems. Int. J. Applied Math. Comput. Sci., 16 (4): 503-511.

Lei, Y. and H. Liu, 2004. Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learning Res., 5: 1205-1224.

Mark, A.H. and A.S. Lloyd, 1998. Feature selection for machine learning: Comparing a correlation_based filter approach to the Wrapper Copyright at American Association for Artificial Intelligence. www.\aaai.org. All Rights Reserved.

Pekalska, E., A. Harol, C. Lai and R.P.W. Duin, 2005. Pairwise selection of features and prototypes. In: Computer Recognition Systems-Proceeding 4th International Conference Computer Recognition Systems, CORES, Advances in Soft Computing, Berlin: Springer, pp: 271-278.

Pudil, P., J. Novovicova and J. Kittler, 1994. Floating search methods in feature selection. Pattern Recogn. Lett., 15 (11): 1119-1125.