A Kernel Based Feature Selection Method Used in the Diagnosis of Wisconsin Breast Cancer Dataset

P. Jaganathan, IEEE member[1], N. Rajkumar[2], and R. Nagalakshmi[3]

[1] Professor and Head, [2]Associate Professor, [3] Student
Dept. of Computer Applications, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India
jaganathodc@yahoo.com, {rknpsna,naga.dharsika}@gmail.com

Abstract. In this paper, a novel feature selection method called kernel F-score is applied for Breast cancer diagnosis. In this method, feature selection for removing the irrelevant/redundant features is achieved in high dimensional spaces than the original spaces. Basically, the datasets in the input space are moved to high dimensional kernel spaces for clear separation of nonlinearity through kernel functions. Then the F-score values for all the features in the kernel space are computed and mean kernel F-score value is set as the threshold for selection or rejection of features. The features lesser than the threshold are removed from feature space. The features above and equal to the threshold are selected for classification and used in the classification of benign and malignant cases using Support Vector Machines (SVM). The results obtained from Wisconsin Breast Cancer Dataset (WBCD) have been satisfied as it produced efficient results than F-score. So, we conclude kernel F-score with SVM for WBCD is promising than F-score with SVM.

Keywords: Feature Selection, Kernel F-score Support Vector Machines, RBF kernel.

1 Introduction

Feature selection process is a technique in data mining widely used in classification tasks. The presence or absence of a feature in any case determines the performance of the classifier in terms of time and cost [2]. Eventhough we have filter and wrapper methods for feature selection, these methods individually produce only fair results when the features are non linear. So it becomes very difficult to select them in the low dimensional space. Therefore kernels are used for nonlinear features separation. Here in this case, features have to be transferred to highdimensional space, where they are comfortably separated.

In order to map the features to high dimensional space Kernel methods are introduced. Kernels select the most discriminative and informative features for classification and data analysis [3]. There are several kernel methods like Kernel Principle Component Analysis (KPCA) has been proposed to obtain non-linear

A. Abraham et al. (Eds.): ACC 2011, Part I, CCIS 190, pp. 683–690, 2011.

principal components [9]. Here in our work we have used radial basis function kernel for mapping and kernel F-score for Wisconsin breast cancer dataset classification.

2 Related Work and Literature Survey

Support vector machine is an effective statistical method used in medical diagnosis for pattern recognition machine learning and datamining (cortes and vapnik 1995). In the literature, there are some works related to breast cancer diagnosis. Among these, Mehmet Faith Akay has proposed a feature selection method with F-score and support vector machines reaching a classification accuracy of 99.51% [7]. Polat et al obtained classification accuracy of 98.53%. With neuro and fuzzy techniques nauck et al produced 95.06% of classification. Goodman et al produced three different results with three different methods such as Optimized-LVQ , Big LVQ, AIRS and accuracies 96.70%,96.80%,97.20% respectively[4].

This research work is supported by All India Council for Technical education, New Delhi under Research Promotion scheme. Ref No. 8023/BDR/RID/RPS/17/08/9

Abonyi and Szeifert (2003) using Supervised fuzzy clustering techniques produced a classification accuracy of 95.57%. logarithmic simulated annealing and perceptron algorithm applied by Albrecht obtained 98.80%. Hamilton et al. (1996) using RIAC method obtained 95.00% classification accuracy[5].with LDA technique Ster and Dobnikar (1996) produced a classification accuracy of 96.80%. Pena-Reyes and Sipper (1999) obtained classification accuracy of 97.36% using Fuzzy-GAI method. Setiono (2000) using Neuro-rule 2a technique obtained classification accuracy of 98.10% . With AR and NN Murat karabatak & M.Cevdet Ince produced classification accuracy of 97.40%[8]. T.S.Subashini et al obtained 97.33% classification accuracy using RBFNN and SVM techniques[10]. Polat et al .have proposed a method called Kernel F-score feature selection (KFFS) used as pre-processing step in the classification of medical datasets[6].

3 Feature Selection

The main idea of feature selection is to select an optimal subset of input variables by removing features with little or no predictive information. There are many feature selection methods. In general it contains two methods which are filter and wrapper methods. The filter methods are independent of learning algorithms where as wrapper methods are dependent on learning algorithms. The F-score method and computation of kernel F-score values are described below.

3.1 F-Score

F-score is a simple method which measures the discrimination of two sets of real numbers. Given training vectors x k , k=1,2,….,m, if the number of positive and

negative instances are n + and n – respectively, then the F-score of the ith feature is defined as

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (1)$$

Where are the average of the ith feature of the whole, positive, and negative datasets, respectively; is the ith feature of the kth positive instance, and is the ith feature of the kth negative instance. The numerator denotes the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score is, the more likely this feature is more discriminative Chen and Lin, 2003[3].The flowchart in Fig1. shows the how the classification accuracy for breast cancer is determined. It demonstrates the computation of kernel F-score values which helps in discriminating the relevant and irrelevant features. Firstly the Kernel Fscore of each feature is calculated and the mean f-score value is determined. The features which are above the mean f-score are selected for classification. With the selected features is passed to SVM classifier with tenfold cross validations. The outcome of this procedure has produces efficient results.

4 Support Vector Machines

Support vector machine is a technique for learning in pattern classification and non-linear regression , pioneered by Cortes and Vapnik in 1995, Boser, Guyon, Vapnik in 1992 and modified by Vapnik in 1999[11]. The main idea of a support vector machine is to construct a hyper plane as the decision surface such that there exists maximum margin between any two different categories. Consider a set of training vectors belonging to two linearly separable classes,

$$(x_i, y_i), x_i \epsilon R^n, y_i \epsilon \{+1, -1\}, i = 1,2, \ldots n. \quad (2)$$

where x_i is a n-dimensional input vector and y_i is a label that determines the class of x_i. A separating hyper plane is determined by an orthogonal vector w and a bias b, which identifies the points that satisfy

$$w.x_i + b = 0 \quad (3)$$

The parameters w and b are constrained by

$$\min |w.x_i + b| \geq 1. \quad (4)$$

A hyper plane in canonical form must satisfy the following constraints,

$$y_i.(w.x_i + b) \geq 1, i = 1,2, \ldots n. \quad (5)$$

The hyper plane that optimally separates the data is the one that minimizes

$$\phi(w) = \frac{1}{2}(w.w). \tag{6}$$

Relaxing the constraints of (4) by introducing slack variables £$_I$ >=0, i=1,2,…,n, becomes

$$y_i.(w.x_i + b) \geq 1 - \xi_i, i = 1,2,\ldots n. \tag{7}$$

In this case the optimization problem becomes

$$\Phi(\mathrm{w},\xi) = \frac{1}{2}(\mathrm{w}.\mathrm{w}) + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{n}} \xi_{\mathrm{i}} \tag{8}$$

with a user defined positive finite constant C. The solution for (7), under the constraints of (6), could be obtained in the saddle point of Lagrangian function

$$\mathrm{L}(\mathrm{w},\mathrm{b},\alpha,\xi,\gamma) - \frac{1}{2}(\mathrm{w}.\mathrm{w}) + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{n}} \xi_{\mathrm{i}} - \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}|\gamma_{\mathrm{i}}\,(\mathrm{w}.\mathrm{x}_{\mathrm{i}} + \mathrm{b}) - 1 + \xi_{\mathrm{i}}| - \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} \gamma_{\mathrm{i}}\,\xi_{\mathrm{i}}, \tag{9}$$

where a_i >=0, £$_I$ >=0, i=1,2,..,n are the Lagrange multipliers. The Lagrangian function has to be minimized with respect to w,b, and £$_I$. Classical Lagrangian duality enables primal problem(8), to be transformed into its dual problem, which is easier to solve. The dual problem is given by

$$\max_{\alpha} \quad [\sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}} - \frac{1}{2}\sum_{\mathrm{ij}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\alpha_{\mathrm{j}}\gamma_{\mathrm{i}}\gamma_{\mathrm{j}}\mathrm{K}(\mathrm{x}_{\mathrm{i}},\mathrm{x}_{\mathrm{j}})] \tag{10}$$

with constraints

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}},\gamma_{\mathrm{i}} = 0, \quad 0 \leq \alpha_{\mathrm{i}} \leq \mathrm{C}, \qquad \mathrm{i} = 1,2,\ldots.,\mathrm{n}. \tag{11}$$

This is a quadratic optimization problem that exists a unique solution. As per K T theorem of optimization theory , the optimal solution satisfies

$$\alpha_i[\gamma_i(w.x_i + b) - 1] = 0, i = 1,2,\ldots n. \tag{12}$$

has non-zero Lagrange multipliers if and only if the points x_i satisfy

$$\gamma_i(w.x + b) = 1. \tag{13}$$

These points are termed SV. The hyperplane is determined by the SV, which is a small subset of the training vectors. Hence if α_i^* is the non –zero optimal solution, the classifier function can be expressed as

$$f(x) = \operatorname{sgn}\left\{\sum_{ij=1}^{n} \alpha_i^* \gamma_i (x_i \,.\, x) + \; b^*\right\} \tag{14}$$

Where b* is the solution of (14) for any non-zero α_i^*.

By defining a non-linear boundary, the SVM constructs an optimal hyperplane in this higher dimensional space. usually non-linear mapping is defined as

$$\phi(.): R^n \rightarrow R^n. \tag{15}$$

In this case, optimal function becomes (15) with the constraints

$$\max_{\alpha} \; [\sum_{i=1}^{n} \alpha_i \; - \frac{1}{2} \sum_{ij=1}^{n} \alpha_i \alpha_j \gamma_i \gamma_j K(x_i, x_j)], \tag{16}$$

Where $K(x_i\ x_j) = \{\varphi(x_i).\varphi(x_j)\}$ is the kernel function performing the non-linear mapping into feature space. The kernel function may be any of the symmetric functions that satisfy the Mercel conditions (Courant &Hilbert, 1953). The most commonly used functions are the Radial Basis Function (RBF): $K(x_i, x_j) = \exp\{-\gamma| x_i - x_j|^2\}$ and the polynomial Function $K(x_i, x_j) = (x_i x_j + 1)^q$, q = 1,2,….,

5 Experimental Observations

Wisconsin Breast cancer dataset:

This database is taken from the UCI machine learning repository for our experiments. It is collected by Dr.William H. Wolberg(1989-91) at the University of Wisconsin-Madison Hospitals.There are 699 records in this database. Each record in the database has nine attributes. The aim of the dataset is to classify the presence or absence of breast cancer given the results of various medical tests carried out on a patient. This database includes 9 attributes. These features are (1) Clump thickness,(2) Uniformity of cell size, (3)Uniformity of cell shape, (4) Marginal adhesion, (5) single epithelial cell cell size, (6) Bare nuclei, (7) Bland chromatin, (8) Normal nucleoli, (9) Mitosis. The nine attributes are represented as an integer value between 1 -10 and detailed in Table 1. In this database, Two hundred and forty one records (65.5%) are malignant and four hundred and fifty eight records (34.5%) are benign[1].In order to evaluate the efficiency of the method, performance measures like sensitivity, specificity, ROC curves, positive predictive value, negative predictive value were considered. The measures were compiled by the following units.

Classification accuracy (%) = TP + TN / TP + FP + FN + TN,

Sensitivity (%) = TP/ TP + FN * 100,

Specificity (%) = TN/ FP + TN * 100,

ROC Curve provides trade-off between sensitivity and specificity.

6 Results and Discussion

In this paper, a new feature selection method called kernel F-score is applied for Wisconsin breast cancer dataset diagnosis. The selected features by applying kernel F-score have been used in the classification of benign and malignant cases using support vector machines. Table 1 shows the obtained reduced number of features before and after applying kernel mapping. we have used two different feature selection methods i) F-Score feature selection without kernel mapping and ii) Kernel F-Score feature selection.Table2 shows the performance of the classifiers with two feature selection methods. Sensitivity, Specificity, Classification accuracy and AUC has been presented. Table3 shows the performance comparison of various training-test partitions with two different methods. 95.70% for 50-50% training-test partition, 95.35% for 60-40% training-test partition, 95.23% for 70-30% training-test partition, 96.41% for 80-20% training-test partition for F-Score with SVM. 96.56% for 50-50% training-test partition, 96.07% for 60-40% training-test partition, 95.71% for 70-30% training-test partition, 96.42% for 80-20% training-test partition for Kernel F-Score with SVM. Fig2 describes ROC curve for kernel F-score with SVM. The results here depicts that our new method Kernel F-Score with Support vector machines for diagnosis of breast cancer produces far better result than F-score combined with Support vector machines.

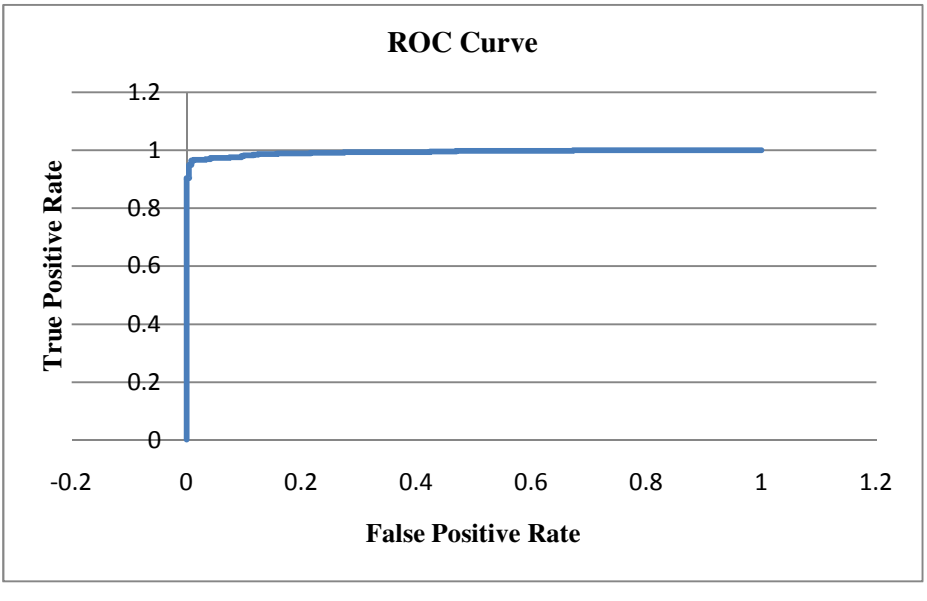

Fig. 2. Area under ROC Curve graph from kernel F-score with SVM

Table 1. The obtained features from kernel mapping

Method	The number of original features in input space	The number of features in kernel space	The number of reduced features with feature selection
F-Score	9	Nil	4
Kernel F-score	9	699	315

Table 2. Performance of the classifier with different methods using Ten-Fold cross validation

Method	Classification accuracy	Sensitivity	Specificity	AUC
F-Score + SVM	95.56	0.97	0.95	0.94
Kernel F-score + SVM	96.99	0.97	0.96	0.97

Table 3. Performance comparison of various training – test partitions with different methods

Method	50-50% training – test partition	60-40% training – test partition	70-30% training – test partition	80-20% training – test partition
F-Score + SVM	95.70	95.35	95.23	96.41
Kernel F-score + SVM	96.56	96.07	95.71	96.42

7 Conclusion

Feature selection is the best technique for obtaining improved classification accuracies in machine learning and pattern recognition. The main idea of feature selection is to select an optimal subset of input variables by removing features with little or no predictive information. In this article kernel F-score feature selection method has been applied for Wisconsin breast cancer dataset diagnosis. In this study, Kernel F-score combined with support vector machine produced better results than F-score method without kernel mapping. The performance measure criteria are classification accuracy, sensitivity–specificity values, and Area under ROC curve values (AUC).The AUC values obtained from F-score and Kernel F-Score with SVM on the classification of Wisconsin breast cancer dataset is found to be 0.94–0.97, respectively. In this way, a new feature selection method is applied on the

classification of WBCD datasets. In future, this method can be applied to other medical datasets which can be used to improve the accuracies in medical diagnosis.

References

[1] Blake, C.L., Merz, C.J.: UCI repository of machine learning database. University of California, Irvine (1998), http://www.ics.uci.edu/mlearn/MLRepository.html

[2] Cao, B., Shen, D., Sun, J.-T., Yang, Q., Chen, Z.: Feature selection in a kernel space. In: International Conference on Machine Learning (ICML), Oregon, USA, June 20-24, pp. 121–128 (2007)

[3] Chen, Y.-W., Lin, C.-J.: Combining SVMs with various feature selection strategies, NIPS 2003 feature selection challenge, 1–10 (2003)

[4] Goodman, D.E., Boggess, L., Watkins, A.: Artificial immune system classification of multiple-class problems. In: Proceedings of the Artificial Neural Networks in Engineering, pp. 179–183

[5] Hamilton, H.J., Shan, N., Cercone, N.: RIAC: A rule induction algorithm based on approximate classification. Technical Report CS 96-06, University of Regina (1996)

[6] Polat, K., Güneş, S.: A new feature selection method on classification of medical datasets: Kernel F-score feature selection. Expert Systems with Applications 36(7), 10367–10373 (2009)

[7] Akay, M.F.: Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications 36(2), Part 2, 3240–3247 (2009)

[8] Karabatak, M., Cevdet Ince, M.: Expert Systems with Applications 36(2), Part 2, 3465–3469 (2009)

[9] Scholkopf, B., Smola, A.J.: Learning with Kernels. The MIT Press, Cambridge (2002)

[10] Subashini, T.S., Ramalingam, V., Palanivel, S.: Expert Systems with Applications 36(3), Part 1, 5284–5290 (2009)

[11] Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995), http://www.emeraldinsight.com