Arif Index for Predicting the Classification Accuracy of Features and its Application in Heart Beat Classification Problem

M. Arif[1], Fayyaz A. Afsar[2], M.U. Akram[2], and A. Fida[3]

[1] Department of Electrical Engineering, Air University, Islamabad, Pakistan
arif@mail.au.edu.pk
[2] Department of Computer and Information Sciences, PIEAS, Islamabad, Pakistan
[3] Department of Electrical Engineering, COMSATS Institute of Information Technology, Pakistan

Abstract. In this paper, Arif Index is proposed that can be used to assess the discrimination power of features in pattern classification problems. Discrimination power of features play an important role in the classification accuracy of a particular classifier applied to the pattern classification problem. Optimizing the performance of a classifier requires a prior knowledge of maximum achievable accuracy in pattern classification using a particular set of features. Moreover, it is also desirable to know that this set of features is separable by a decision boundary of any arbitrary complexity or not. Proposed index varies linearly with the overlap of features of different classes in the feature space and hence can be used in predicting the classification accuracy of the features that can be achieved by some optimal classifier. Using synthetic data, it is shown that the predicted accuracy and Arif index are very strongly correlated with each other ($R^2 = 0.99$). Implementation of the index is simple and time efficient. Index was tested on Arrhythmia beat classification problem and predicted accuracy was found to be in consistent with the reported results.

Keywords: Clustering, Pattern Classification, Features, Nearest Neighbor Search, Classification Accuracy.

1 Introduction

In pattern classification problem, classification accuracy depends on proper selection of features that can discriminate different classes and design of a good classifier. Design of a classifier includes ability of classifier to approximate decision boundary of arbitrary complexity among different classes in the feature space and generalization power of the classifier. Discrimination power of features decides maximum possible classification accuracy achievable by any classifier. A prior knowledge of maximum achievable classification accuracy can help a lot in designing appropriate optimal classifier. Moreover, if classes are separable in the feature space by a decision boundary of any arbitrary complexity, then it is possible to achieve maximum classification accuracy. In real life, feature representing different classes can cluster in the feature

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 491–498, 2009.

space as scattered multi-modal clusters. A class can be represented by multiple clusters scattered in the feature space. These clusters can be point or line or any arbitrary shaped clusters. Moreover, they can be compact and well separated (separated by large margin) within class or overlapping each other. Overlapping of intra-class clusters has less effect on classification accuracy as compared to overlap among inter-class clusters. Hence, for a good feature classification, clusters in a particular class need not be very compact and well separated in the feature space but the decision boundary among classes should be well separated and non-overlapping. Hence an index is required to assess the discrimination power or quality of the features in the feature space when information of class labels is available. This index should posses following characteristics. It should be sensitive to the amount of overlap of features among different classes that results in the decrease of classification accuracy and strong correlation may exist between index value and amount of overlap or classification accuracy. If classes in the feature space are separable by any decision boundary of arbitrary complexity, it should give a consistent value and the value of index may not vary with the complexity of decision boundary. Index may not vary with the number of clusters per class, shape of the clusters and their intra-class overlap and their location with respect to other classes in the feature space. Such an index can give us a prior knowledge of maximum achievable classification accuracy by any perfect classifier.

Three different kinds of clustering validity indices exist in the literature, namely, external criteria based indices, internal criteria based indices and relative criteria based indices [2]. In internal criteria based indices, clusters generated from a clustering algorithm are evaluated by considering the data itself based on intra-cluster and inter-cluster distances. In these indices, Dunn's Index [3], Alternative Dunn's Index [1], Davies-Boulden Index [4] and Xie and Beni's Index [5] are worth mentioning. In external criteria based indices, user defined partition of data is provided (class labels) that can be compared with the clustering structure revealed by a clustering algorithm. Different external criteria based indices like Rand Index [6], Jaccard Co-efficient [2], Folkes and Mallows [7], Mirkin Index [8] and Adjusted Rand Index [9] are reported in the literature. Density based clustering algorithms can find arbitrary shaped clusters. Many such algorithms are reported in the literature like DBSCAN [11], BRIDGE [12], DBCLASD [13] and DENCLUE [14]. These indices are not useful in evaluating the discrimination power of the features in classification problem. These indices cannot be applied directly to the feature representation of different classes in pattern classification problem by simply assuming classes as clusters. Clustering structure within class is required before applying these indices. A survey of clustering algorithms can be found in [10]. Performance of these indices is very sensitive to the performance of clustering algorithm. Clustering algorithm should be capable of discovering cluster structure of any arbitrary shape. In this paper, an index called Arif index is proposed which does not require any clustering algorithm. This index spreads linearly on the scale of zero and one where zero value shows no overlap among clusters of different classes. This index is independent of number of clusters per class, type of clusters and their location in feature space. We have used Arif Index to assess the discrimination power of the features used in Arrhythmias beat classification problem [15].

2 Description of Proposed Arif Index

Let number of classes is N_C having data points $n_i, i = 1,...,N_c$ and total data points are $N = \sum_{i=1}^{N_C} n_i$. Dimension of the feature space is *d*. Algorithm of Arif index is described as below,

Step 1: Normalize the feature vectors by making their means equal to zero and variances equal to one.

Step 2: initialize a variable Status S_t of size $N \times 1$ with zeroes.

Step 3: For a data point *y* of size $d \times 1$ of j^{th} Class, Find nearest neighbor of y in the rest of classes which are different from j^{th} Class. Let this nearest neighbor is nn_d having distance δ_{y,nn_d} from data point *y*.

Step 4: Find all data points of j^{th} Class whose distance are less than δ_{y,nn_d} from data point *y*. Let number of nearest neighbors of y in j^{th} Class whose distance are less than δ_{y,nn_d} are $nn_s(k), \quad k = 1,2,...,n_j$. Average number of neighbors near to each other is defined as,

$$C(j) = \frac{1}{n_j} \sum_{k=1}^{n_j} nn_s(k) \tag{1}$$

Step 5: If number of nearest neighbors nn_s is greater than a user defined threshold value γ, consider this data points as clustered in the data points of the same class and make the status of the set of data points $\{y \quad nn_s\}$ equals to 1.

Step 6: Run steps 2 to 5 for all the data points in the feature space.

Step 7: Arif index will be defined as follows,

$$AF_{index} = \frac{N - \sum_{i=1}^{N} S_t(i)}{N} \tag{2}$$

Hence Arif index gives the ratio of data points which are not surrounded by data points of its class to the total number of data points. Strength of clustering data points of the same class near a particular data point is controlled by a user defined threshold value γ. The values of γ should be greater than 1. Value of Arif index varies from 0 to 1, where value of 0 means no overlapping and maximum accuracy of 100% is achievable and values of 1 means complete overlap and accuracy depends on the data representation of different classes. In case of two classes with equal number of

representation, for Arif index equals to 1 means 50% accuracy is possible by just assigning label of one class to all the data. $C(j)$ is the average number of nearest neighbor of all the feature vectors of j^{th} class. This value will give the density estimate of the clustering structure of a particular class. Low value of $C(j)$ shows sparse representation of a class in the feature space. Hence, it will also help in better understanding the quality of features representing the pattern classification problem in the feature space.

3 Results and Discussion

Checkerboard data as shown in Figure 1 is used to evaluate the Arif index for separable classes. Furthermore, data of two classes are brought together in steps so that overlapping of both classes increases and at the end they completely overlap each other. K-nearest neighbor is used as a classifier and value of K is set to 5. Half of the data is used for the training and rest of the data is used for the testing purpose. Arif index is calculated on training data and classification accuracy of K-nearest neighbor classifier is calculated on testing data. In Figure 2, scatter plot between Arif index and the classification accuracy is plotted. It can be observed from the Figure 2 that when two classes are separable, the value of Arif index was zero and as overlapping between classes increased, the value of Arif index also increased and the classification accuracy decreased. Arif index approached near to its maximum value of 1 for completely overlapping classes. A very nice agreement between values of Arif index and classification accuracy is observed. A linear trend is very prominent in the scatter plot and straight line is fitted with very high value of $R^2 = 0.99$. For the value of Arif index near to 1, accuracy dropped to almost 50%. This observation is very obvious as we have used equal number of data points for each class in the testing data. If we do not apply any classifier and assign one class label to all the data points of the testing data, we will get the accuracy of 50% without applying any classifier. This shows the usefulness of using Arif index to get an idea about the upper limit of maximum achievable accuracy before applying any classifier to the data.

For multi-class problems, lower bound on the accuracy without applying any classifier will be the percentage of representation of majority class in the data set if we set all the class labels to the class label of majority class. A linear trend can be interpolated between 100% classification accuracy and the lower bound of the classification accuracy. An ideal linear trend can be described as below,

$$Accuracy = 100 - (100 - Accuracy_{lower\_bound}) \times Arif\ \ Index \qquad (3)$$

Where $Accuracy_{lower\_bound}$ is the lower bound on the accuracy depending on the percentage representation of the majority class. To prove this hypothesis, we have generated data for four separable classes as shown in Figure 3a and slowly move them towards each other until they completely overlap each other as shown in Figure 3b. During this process, we have calculated Arif index and accuracy by applying a K-nearest neighbor classifier and plotted it in Figure 4. It can be observed from Figure 4 that for Arif index equals to 1 (case of completely overlap), accuracy dropped to 25%.

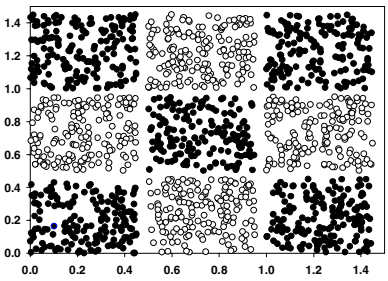

Fig. 1. Checkerboard data of two classes

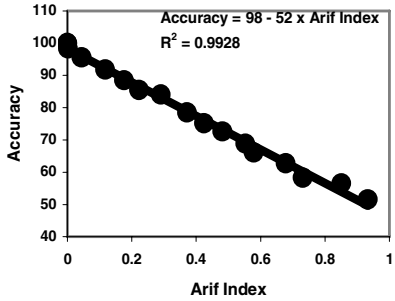

Fig. 2. Plot of Classification Accuracy and Arif Index for all five data sets

A linear trend can be fitted on these data points with very high value of regression coefficient $R^2 = 0.98$. Since there were four classes containing equal number of data points, the classification accuracy at complete overlap will be 25%. This supports our hypothesis and applicability of equation (3). Hence Arif index can be used to predict the quality of features before applying any classifier and value of Arif index can be used to predict maximum achievable classification accuracy that can help in proper tuning or selection of the classifiers.

ECG beat classification, being an integral part of any ECG based automatic decision support system, has been studied by a number of researchers. Different feature extraction methods for beat classification include use of Fourier Transform [16], multi-resolution analysis [17], wavelet transform [18] etc. In our previous work [15], we have used features extracted from two-level wavelet decomposition of an ECG signal from MIT-BIH Arrhythmia Database [19]. It contains two-channel ambulatory ECG recordings from 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979. ECG recordings were digitized at the sampling rate of 360 Hz with 11-bit resolution. We have used the annotations of the cardiologist originally provided by the database. In this paper, same set of features are used as in [15] to predict the classification accuracy of the beat classification using Arif index. Eleven features extracted from the wavelet decomposition of the ECG beat signal and RR interval are used for classification. Further Principal Component Analysis is used to reduce the

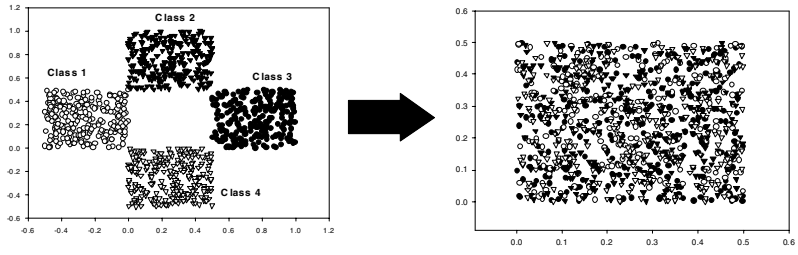

Fig. 3. (a) Set 2: Four class data having four separable square clusters (b) Complete overlap of all four classes

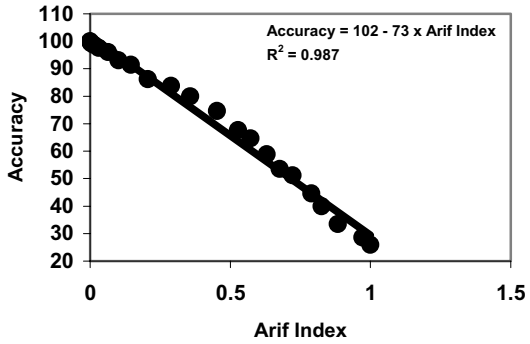

Fig. 4. Plot of Classification Accuracy and Arif Index for Set 6

Table 1. PPV and C values for Set 6

Class	PB	APB	LBBB	N	RBBB	PVC
Data #	2400	1700	4800	7200	4800	2300
PPV (Without PCA)	99.82	99.00	99.24	99.16	99.11	99.95
C	890	550	526	2874	513	337
PPV (With PCA)	99.78	99.17	99.27	98.80	98.85	99.89
C	934	524	645	2656	508	350

dimension of the feature set from eleven to six. For details of the feature extraction, please refer to [15]. Two sets, Set6 and Set9, are constructed from the database. Set6 consists of 23200 beats of six types (Paced Beats (PB), Atrial Premature Beat (APB), Premature Ventricular Contraction (PVC), Normal (N), Left and Right Bundle Branch Blocks (LBBB & RBBB)), whereas Set9 consists of 104690 beats of nine types (PB, APB, PVC, N, LBBB, RBBB, Fusion of paced and Normal beats (P&N), Fusion of Normal and Ventricular beats (N&V) and Ventricular Flutter (VF)). A simple *K*-nearest neighbor classifier has been employed for the classification of different types of beats. Arif index along with *C* were calculated on Set6 and Set9. Classification accuracies were calculated by using half of the data sets for training and rest of the half for testing.

Positive Predictive Values (*PPV*) of each class was calculated as follows,

$$PPV_c = \frac{TP_c}{TP_c + FP_c} \tag{4}$$

Where TP_c and FP_c representing the number of true and false positives for a given class *c*. PPV for Set6 is tabulated in Table1 with the values of C obtained from Arif index as explained in Section 2. Values of *C* represent number of neighbors of a data point of same class within the distance of the data point to the data point of other class. Hence high value of *C* shows denser clustering and low value corresponds to

Table 2. Arif Index, predicted and achieved accuracies for Set 6

	Arif Index	Predicted Accuracy	Achieved Accuracy
Without PCA	0.004	99.7	99.49 [15]
With PCA	0.004	99.7	99.47 [15]

Table 3. PPV and C values for Set 9

Class	VF	PB	APB	N&V	LBBB	N	RBBB	PVC	P&N
Data #	470	3616	2495	774	8067	74716	7247	7058	258
PPV (Without PCA)	80.48	99.66	83.54	77.68	94.46	98.27	98.16	92.6	88.07
C	7	737	73	11	176	1100	424	143	10
PPV (With PCA)	79.51	99.43	82.51	77.08	93.95	98.15	98.00	92.3	78.48
C	7.8	624	62	11	188	1041	379	133	7.8

Table 4. Arif Index, predicted and achieved accuracies for Set 9

	Arif Index	Predicted Accuracy	Reported Accuracy
Without PCA	0.023	99.3	97.04 [20]
With PCA	0.068	98	96.82 [20]

sparse representation of a class. Arif index, predicted accuracy and achieved accuracy are given in Table 2. It can be seen from the table that Arif index for Set6 is low; nearly zero, and predicted accuracies and achieved accuracies match each other. *C* values tabulated in Table 1 are high showing denser clustering and hence PPV values for all six classes with or without PCA are very high, i.e. above 99%. For Set9 with larger number of beat types, results are tabulated in Table 3 and 4. It can be observed from Table 4 that for low values of C (VF, N&V and APB), PPV values are also low as compared to denser classes. For the reduced feature sets (With PCA), values of *C* further reduced and this reduction is reflected in the decrease of PPV values of less denser classes. Arif index, predicted accuracy and achieved accuracy for Set9 are tabulated in Table 4. Arif index is increased a bit as compared to Set6 and correspondingly predicted and achieved accuracies are also dropped slightly. Hence it is proved that Arif index and C values can be used efficiently to predict the classification accuracy and PPV of individual classes.

4 Conclusions

A novel index to assess the classification quality of the feature vectors is proposed in this paper. This index is a model free index and does not require any clustering algorithm. It only uses the information of local neighborhood of the feature vectors to calculate the overlap or region of confusions in the feature space. Results have shown that value of the proposed index does not depend on the shape, location or the structure of the clusters in the feature space. Moreover, values of the Arif index are shown to be strongly correlated with the classification accuracies. Predicted accuracies of different physiological sets are also found to be consistent with the reported accuracies in the literature. Hence this index will be very useful for the pattern classification problems.

References

1. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On Clustering Validation Techniques. Journal of Intelligent Information Systems 17(2/3), 107–145 (2001)
2. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster validity methods: Part 1. In: SIGMOD Record, vol. 31(2), pp. 40–45 (2002)
3. Dunn, J.C.: Well Separated Clusters and Optimal Fuzzy Partitions. J. Cybern. 4, 95–104 (1974)
4. Davies, D.L., Bouldin, D.W.: A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1(2), 224–227 (1979)
5. Xie, X.L., Beni, G.: A Validity Measure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(4), 841–846 (1991)
6. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66, 846–850 (1971)
7. Fowlkes, E., Mallows, C.: A method for comparing two hierarchical clustering. Journal of the American Association 78 (1983)
8. Mirkin, B.G., Cherny, L.B.: On a distance measure between partitions of a finite set. Automation and remote Control 31(5), 91–98 (1970)
9. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification, 193–218 (1985)
10. Xu, R., Wunsch, D.: Survey of Clustering Algorithms. IEEE Transactions on Neural Networks 16(3), 645–678 (2005)
11. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231 (1996)
12. Dash, M., Liu, H., Xu, X.: 1+1>2: Merging Distance and Density Based Clustering. In: Proceedings of Seventh International Conference on Database Systems for Advanced Applications, pp. 32–39 (2001)
13. Xu Ester, X., Kriegel, M., Sander, H.-P.: A distribution-based clustering algorithm for mining in large spatial databases. In: Proceedings of 14th International Conference on Data Engineering, pp. 324–331 (1998)
14. Hinneburg, A., Keim, D.A.: An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp. 58–65 (1998)
15. Afsar, F.A., Arif, M.: Robust electrocardiogram (ECG) beat classification using discrete wavelet transform. Physiological Measurement 29, 555–570 (2008)
16. Minami, K., Nakajima, H., Toyoshima, T.: Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network. IEEE Transactions on Biomedical Engineering 46(2), 179–185 (1999)
17. Prasad, G.K., Sahambi, J.S.: Classification of ECG arrhythmias using multiresolution analysis and Neural Networks. In: Conference on Convergent Technologies, India (2003)
18. Yu, S.N., Chen, Y.H.: Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network. Pattern Recognition Letters 28(10), 1142–1150 (2007)
19. Mark, R., Moody, G.: MIT-BIH Arrhythmia Database Directory. MIT Press, Cambridge (1988)
20. Usman Akram, M.: Application of Prototype Based Fuzzy Classifiers for ECG based Cardiac Arrhythmia Recognition, BS Thesis, Pakistan Institute of Engineering and Applied Sciences (2008)