Evolving Linear Discriminant in a Continuously Growing Dimensional Space for Incremental Attribute Learning

Ting Wang[1,2*], Sheng-Uei Guan[2], T. O. Ting[3], K. L. Man[2], Fei Liu[4]

[1] Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK
[2] Department of Computer Science and Software Engineering,
[3]Department of Electrical and Electronic Engineering,
Xi'an Jiaotong-Liverpool University, Suzhou 215123, P.R. China
[4] Department of Computer Science & Computer Engineering, La Trobe University, Victoria 3086, Australia

ting.wang@liverpool.ac.uk, steven.guan@xjtlu.edu.cn, toting@xjtlu.edu.cn, kalok2006@gmail.com, f.liu@latrobe.edu.au

Abstract. Feature Ordering is a unique preprocessing step in Incremental Attribute Learning (IAL), where features are gradually trained one after another. In previous studies, feature ordering derived based upon each individual feature's contribution is time-consuming. This study attempts to develop an efficient feature ordering algorithm by some evolutionary approaches. The feature ordering algorithm presented in this paper is based on a criterion of maximum mean of feature discriminability. Experimental results derived by ITID, a neural IAL algorithm, show that such a feature ordering algorithm has a higher probability to obtain the lowest classification error rate with datasets from UCI Machine Learning Repository.

Keywords: pattern classification, incremental attribute learning, data preprocessing, feature ordering, neural networks

1 Introduction

Incremental Attribute Learning (IAL) is a machine learning strategy where features in the problem are often gradually trained one by one. Such a machine learning approach is usually employed to solve complex pattern recognition problems. During the solution process, features with greater discriminability are distinguished and separated from features with weaker discriminability by some criteria in the first place. After that, some approaches like neural networks and genetic algorithms can be employed to complete the incremental training. Therefore, there are two important steps in the processing. One is the criterion for the differentiation of features with great discriminability; while the other is the machine learning approach for pattern recognition. In the first step, features are sorted in some order from high discriminability to low dis-

*Corresponding Author

criminability. In the second step, ordered features are trained by incremental attribute machine learning approaches for classification or regression. Consequently, the criterion of feature ordering is regarded as the key to enhancing performance of final results from incremental machine learning.

In previous research, feature ordering criteria have been developed based on some feature selection approaches [1-3]. Generally, these criteria can be divided into two types: wrappers and filters. The former is based on each feature's individual contribution to classes; while the latter is based on the score of each feature according to some ranking mechanisms like Linear Discriminant or Correlation.

In this study, an evolutionary feature ordering criterion is presented. According to this criterion, feature ordering preprocess of IAL aims to use an evolutionary algorithm to search the training sequence of features where the mean of features' discriminabilities calculated in this sequence is the maximum one compared with the other feature orderings. In this paper, IAL and feature ordering are reviewed in section 2. An Accumulative Linear Discriminant for the calculation of feature discriminability is presented in section 3. Section 4 illustrates the maximum discriminability mean criterion, and an evolutionary algorithm of the criterion is interpreted in section 5. Experimental results and analysis for benchmarks from UCI machine learning dataset are illustrated in section 6. Finally, section 7 concludes this study with key findings.

2 Incremental Attribute Learning with Ordered Features

2.1 Feature Ordering in IAL

Incremental Attribute Learning is a “divide-and-conquer” machine learning strategy where features are gradually trained in stages. Compared with Incremental Learning (IL) where the number of training patterns is gradually increasing, IAL focuses on the increasing of the feature number in a machine learning process. IAL aims to solve easy problems earlier and cope with difficulties later. Because of the segmentations of features, IAL can avoid the curse of dimensionality in high-dimensional problems. It is also applicable for problems with newly imported features.

Previous studies showed that IAL can improve final performance in pattern recognition. Particularly, IAL can bring along more accurate results than conventional approaches where features are imported to training by batch. For example, based on UCI datasets, classification errors of Diabetes, Thyroid and Glass derived by ILIA [4] and ITID[1], two neural IAL algorithms, reduced by 8.2%, 14.6% and 12.6%, respectively [1, 2]; moreover, based on OIGA, testing error rates derived by IGA of Yeast, Glass and Wine declined by 25.9%, 19.4% and 10.8% [5] in classification. Furthermore, i^+Learning and i^+LRA, two kinds of IAL decision trees, were employed to run 16 different UCI datasets. Results indicated that algorithms based on IAL can get better performance than ITI in 14 datasets of the total [6]. In addition, a study on incremental SVM extended IAL to a wider application field [7]. All of these previous IAL studies showed that IAL can indeed promote the performance of pattern recognition.

Moreover, in previous studies, the significance of feature ordering to improving final results in pattern recognition [2, 5] has been discovered. Feature ordering is sel-

dom used in conventional methods where features are trained in one batch, in contrast, feature ordering affects the training results from IAL. Thus, feature ordering is unique to IAL. Previous studies sorts feature ordering in two different ways: ranking-based filters and contribution-based wrappers. Such a division is similar to those approaches in feature selection. Previous studies have validated that ranking-based feature ordering approaches are better than the contribution-based ones usually at least in two different aspects: time [8] and error rate [9]. Different from feature selection, which attempts to search a feature subset or reduce feature weights for the optimal results, feature ordering aims to sort features for IAL purpose by some criteria. It is obvious that different criteria produce different feature ordering that may produce different results.

2.2 Incremental Neural Networks

ITID [1], a representative of neural IAL based on ILIA [4], is different from traditional approaches which train features by batch. It divides all input dimensions into several sub-dimensions, each of which corresponds to an input feature. Instead of learning input features altogether as an input vector in training, ITID learns inputs through their corresponding sub-networks one after another and the structure of neural networks gradually grows with an increasing input dimension as shown in Fig. 1. During the training, information obtained by a new sub-network is merged together with the information obtained by the old network. ITID has a pruning technique which is adopted to find the appropriate network architecture. With less internal interference among input features, ITID achieves higher generalization accuracy than conventional methods [1].

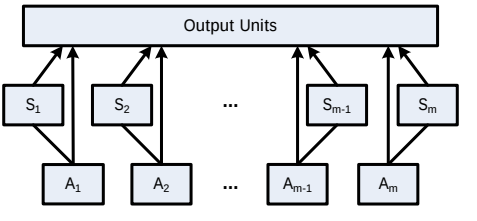

Fig.1. The basic network structure of ITID [1].

3 Linear Discriminant for Feature Ordering

Linear discriminant is usually employed to evaluate feature discriminability in the dimensional space where feature number is stable. However, feature dimension in IAL is dynamic, thus conventional linear discriminant should be adapted.

3.1 Fisher's Linear Discriminant

FLD, a linear statistical classifier, provides simple ways to estimate the accuracy of classification problems. It firstly assumes that the datasets used in FLD are Gaussian conditional density models, where data have normal distributed classes or equal class covariance. The Fisher criterion aims to search a direction where the distance between different classes is the farthest and the distance of each pattern within every class is the closest. Thus, in this direction, the ratio of distance between-classes and within-classes is the largest compared with other directions. Such a direction often leads to the simplest classification. Mathematically, FLD in two-category classification is

$$J(w) = \frac{(\tilde{\mu}_2 - \tilde{\mu}_1)^2}{s_1^2 + s_2^2} \tag{1}$$

where$\tilde{\mu}_1$ and $\tilde{\mu}_2$are two means of projected classes, and s_1 and s_2 are within-class variances. The objective of FLD is to search the matrix *w* for maximum *J*(*w*). The larger the *J*(*w*), the easier in the classification.

However, due to the fact that *J*(*w*) is impacted by two classes, it will be difficult to calculate *J*(*w*) for patterns belonging to three or more classes at a same time. Therefore, (1) should be revised for this demand.

3.2 Standard Deviation Linear Discriminant in IAL

In IAL, each feature's discriminability can be estimated in this feature's one-dimensional space. Features can be ordered by the ranking value of feature discriminability. For two-class classification problems c_2, based on formula (1), discriminability of feature f_i can be given by

$$D(f_i) = \frac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2} \tag{2}$$

where μ_1 and μ_2 are two means of classes, and s_1 and s_2 are within-class variances.

However, (2) is too simple to cope with multi-category classifications, because the between-class scatter is difficult to describe merely by distance between patterns. Here, the difference between the centers of these multiple classes should be replaced by standard deviations of centers and standard deviations of patterns, so that the influence brought by classes whose mean is not the smallest or the largest among all means of classes can be measured.

Definition 1. Single Discriminability (SD) is a ratio of a feature by the standard deviation of all class centers and the sum of standard deviations of all patterns in each class.

SD for both two-category and *n*-category classification problems can be unified as

$$D(f_i) = \frac{std\left[\left(\mu_{f_i,j}\right)_{j=1}^{j=n}\right]}{\sum_{j=1}^{j=n} std\,(f_i)_j} \tag{3}$$

where *n* is the total number of classes, and two *std*s are standard deviations, one for all patterns belonging to c_j in feature *i*, and the other for the vector consisting of the

means of all classes. Let **x** be the vector for standard deviation calculation, the standard deviation of **x** is

$$std(\mathbf{x}) = \sqrt{\frac{\sum_{k=1}^{k=r}(x_k - \mu)^2}{r-1}} \tag{4}$$

where the vector$\mathbf{x} = \{x_k\}_{k=1}^{k=r}$,$x_k$ is the value of k^{th} pattern, and r is the total number of patterns. Obviously, in equation (4), the part of $(x_k - \mu)$ is a distance between k^{th} pattern and its mean. Thus, let *dist* replace this part, (4) can be re-written as:

$$std(\mathbf{x}) = \sqrt{\frac{\sum_{k=1}^{k=r} dist_{x_k,\mu}{}^2}{r-1}} \tag{5}$$

where$\mathrm{dist}_{x_k,\mu}$denotes the distance of k^{th} pattern in x and its mean μ.

3.3 Linear Discriminant for a Growing Feature Space

With the increasing number of new features in IAL, the dimension number of feature space is also growing. A growing feature space has been regarded as one of the most manifest characteristics of IAL. In such a growing feature space, the standard deviation, which is the core of evolving linear discriminant, should be upgraded from that in one dimension. More specifically, the standard deviation in one dimensional space is based on the distance between each pattern and their mean of the same class. This distance should be extended to a higher dimensional space, when the feature space is growing. If ‖D‖ is the Euclidean norm of *d*-dimensional feature space, (5) can be given in a high-dimensional style by

$$std(\mathbf{x}) = \sqrt{\frac{\sum_{k=1}^{k=r}\left\|D_{x_k,\tilde{\boldsymbol{\mu}}}\right\|^2}{r-1}} \tag{6}$$

where$\tilde{\mu}$ is the barycenter of x, and

$$\left\|D_{x_k,\tilde{\boldsymbol{\mu}}}\right\| = \sqrt{\sum_{i=1}^{d}\left(x_{k,i} - \mu_i\right)^2} \tag{7}$$

where d is the total number of features imported so far. Therefore, to calculate standard deviation of r patterns in two dimensions, (6) can be written as

$$std(\mathbf{x}) = \sqrt{\frac{\sum_{k=1}^{k=r}\left[\left(x_{k,1} - \mu_1\right)^2 + \left(x_{k,2} - \mu_2\right)^2\right]}{r-1}}, \quad \mathbf{x} = \left\{x_{k,d}\right\}_{k=1,d=1}^{k=r,d=2} \in \mathbb{R}_{feature}^{r\times 2} \tag{8},$$

and for a tri-dimensional space, the equation is

$$std(\mathbf{x}) = \sqrt{\frac{\sum_{k=1}^{k=r}\left[\left(x_{k,1} - \mu_1\right)^2 + \left(x_{k,2} - \mu_2\right)^2 + \left(x_{k,3} - \mu_3\right)^2\right]}{r-1}} \tag{9}.$$

where$\mathbf{x} = \left\{x_{k,d}\right\}_{k=1,d=1}^{k=r,d=3} \in \mathbb{R}_{feature}^{r\times 3}$

Accordingly, multidimensional standard deviation of r patterns in an m-dimensional space is

$$std(\mathbf{x}) = \sqrt{\frac{\sum_{k=1}^{k=r}\sum_{i=1}^{i=m}\left(x_{k,i}-\mu_i\right)^2}{r-1}}, \mathbf{x} = \left\{x_{k,d}\right\}_{k=1,d=1}^{k=r,d=m} \in \mathbb{R}_{feature}^{r\times m} \quad (10).$$

Based on (10), when some new features are incrementally introduced into the system, formula (3), the standard deviation based linear discriminant of IAL in one feature dimension, should be upgraded to fit in this gradually increasing dimensional space, because (3) has little consideration on gradually importing of new features.

Definition 2. Accumulative Discriminability (AD) is the ratio in d-feature space between the multidimensional standard deviation of all class centers and the sum of all multidimensional standard deviations of all patterns in each class.

If $\{f_1, f_2, \ldots, f_m\}$ is the pool of input features, $\mathbf{f} = \left\{f_{k,d}\right\}_{k=1,d=1}^{k=r,d=m} \in \mathbb{R}_{featur\ e}^{r\times m}$, when the d^{th}($1 \le d \le m$) feature is imported, AD is

$$AD(f_1, f_2, \ldots, f_d) = \frac{std\left[\left(\tilde{\boldsymbol{\mu}}_j\right)_{j=1}^{j=n}\right]}{\sum_{j=1}^{j=n} std\left[\left(f_i\right)_{i=1}^{i=d}\right]_j}, (1 \le d \le m) \quad (11)$$

where $\tilde{\boldsymbol{\mu}}_j$ is the barycenter of vector $(f_1, f_2, \ldots, f_d)$ with patterns belonging to j.

Therefore, results of (11) are dynamic when new features are gradually imported into training. To obtain better classification results, it is necessary to ensure the result of (11) being the maximum in every step of feature importing.

4 Maximum Mean Discriminative Criterion

To obtain the most accurate classification result in IAL, it is necessary to ensure datasets have the greatest discriminability in every step when a new feature is imported into the predictive system and the feature dimension is increased from d to $d+1$. Therefore, the ratio in (11) will be the largest all the time, which guarantees different classes always can be separated in the easiest way. Therefore, the criterion for optimal classification results, also the greatest discriminability, is to produce an optimal feature ordering which contains the greatest discrimination ability in each round of feature importing. Obviously, after all features are imported, the optimal feature ordering will have the largest sum or mean of features' discriminability calculated in each step of the process. Hence such a criterion for obtaining the optimal feature ordering can be given with maximum discriminability mean by

$$\max \frac{1}{d}\sum_{d=1}^{d} AD(\mathbf{f}_{1:d}), (1 \le d \le m) \quad (12)$$

where $\mathbf{f}_{1:d}$ is the feature subset of $\{f_1, f_2, \ldots, f_m\}$ during the feature importing process.

Usually, feature with greater SD calculated by (3) may not always get the greater AD, because (11) has an additional value produced by the Euclidean distance in high dimensional space. Such a value is disproportionate with the value in (3). Thus features which have greater SD may also have weaker AD in IAL feature importing.

Therefore, for IAL classification, (12) will likely produce more accurate results than (3).

5 Evolution of the Optimal Feature Ordering

An evolutionary algorithm can be employed to obtain the maximum mean of features' discriminability for optimal feature ordering. An algorithm modified by Genetic Algorithms (GA) is employed.

Firstly, the algorithm randomly produce a set of seeds in different feature orderings. Secondly, more than two places in the ordering of each seed are exchanged to generate a new ordering. Such an exchange is similar to crossover and mutation in GA. According to criterion (12), if the seed gets the greatest mean of discriminability in its history, it will be recorded, and after several epochs of evolution, the recorded feature ordering of each seed will be compared with one another, and a seed with the greatest mean will be selected as global optimization.

In global optimization, sometimes because of the large feature number and limitations of the evolutionary generation number, the selected seed is only a potential global optimal solution, which is close to the real optimal solution. Therefore, this seed should be evolved again to search for the real one (global optimum). Here, only the potential seed is evolving. After a number of evolutions, if there is no better one, the recorded global feature ordering seed will be regarded as the truly optimal one. Therefore, usually there is a limitation to the maximum mean of discriminability.

To guarantee final feature ordering has the maximum mean, it is necessary to repeat global optimization searching process for several times. If most of repetitions produce the same results, and the results have the greatest mean of discriminability, these same results can be concluded as a global optimum.

Obviously, the ordering transformed data based on the global optimal feature ordering can be directly employed in training, validation and testing. The speed of producing such a transformed dataset depends on the feature dimensional numbers, the number of evolving generations and the number of random seeds.

6 Experiments

A summary of four benchmarks, Diabetes, Cancer, Glass, and Thyroid, from UCI Machine Learning Repository in our experiments is show in Table 1.

Table 1.brief information of datasets

	DIABETES	CANCER	GLASS	THYROID
PATTERN	768	699	214	7200
INPUT NUMBER	8	9	9	21
OUTPUT	2	2	6	3

In the experiments, 50% patterns were randomly selected as training data, 25% for validation and 25% for testing. Moreover, for AD optimal feature ordering, 100 seeds were generated in a 10-generation evolution. Potential global optimum feature ordering was obtained and repeated for 10 times to confirm the optimization status. Diabetes, Cancer, and Glass have 200 generations in each confirmation round, while Thyroid has 5000 epochs, because of its large feature number. Results based on ITID and feature orderings derived by AD were compared with those results obtained in previous studies, where feature orderings were derived based on original orderings [2], wrappers [4], correlation-based mRMRs [9, 10], and conventional approaches [4]. Here, ITID was randomly initialized by 20 different structures, and the final results were the statistical average of these 20 different initial neural networks. Table 2-5 show the comparison results of classification error rate and the means of AD. Results derived in this study have been highlighted in bold.

Table 2. Diabetes results comparison

Approaches	Feature Ordering	Classification Error Rate
ITID-AD (AD)	**2-6-7-8-5-4-1-3**	**21.61458%**
ITID-SD (SD)	2-6-8-7-1-4-5-3	21.84896%
mRMR-Diff.(mRMRd)	2-6-1-7-3-8-4-5	22.86459%
mRMR-Quo. (mRMRq)	2-6-1-7-3-8-5-4	22.96876%
Wrappers (Wra.)	2-6-1-7-3-8-5-4	22.96876%
Original (Ori.)	1-2-3-4-5-6-7-8	22.86458%
Conventional (Con.)	In One Batch	23.93229%

Table 3. Cancer results comparison

Approaches	Feature Ordering	Classification Error Rate
ITID-AD	**3-2-6-7-5-1-8-4-9**	**1.551726%**
ITID-SD	3-2-6-7-1-8-4-5-9	1.695405%
mRMR-Difference	2-6-1-7-3-8-5-4-9	2.29885%
mRMR-Quotient	2-6-1-7-8-3-5-4-9	2.29885%
wrappers	2-3-5-8-6-7-4-1-9	2.4999985%
Original Ordering	1-2-3-4-5-6-7-8-9	2.902299%
Conventional	In One Batch	1.867818%

Table 4. Glass results comparison

Approaches	Feature Ordering	Classification Error Rate
ITID-AD	**3-8-2-4-6-7-1-5-9**	**34.33964%**
ITID-SD	3-8-4-2-6-5-9-1-7	34.81133%
mRMR-Difference	3-2-4-5-7-9-8-6-1	39.05663%
mRMR-Quotient	3-5-2-8-9-4-7-6-1	35.28304%
wrappers	4-2-8-3-6-9-1-7-5	36.4151%
Original Ordering	1-2-3-4-5-6-7-8-9	45.1887%
Conventional	In One Batch	41.226405%

Table 5. Thyroid results comparison

APP.	FEATURE ORDERING	ERROR RATE
AD	**21-18-19-15-20-17-13-7-12-5-4-8-3-9-16-6-14-1-11-10-2**	**1.525001%**
SD	21-19-17-18-3-7-6-16-13-20-10-8-2-4-5-1-11-12-14-15-9	1.927777%
MRMRD	3-7-17-10-6-8-13-16-4-5-12-21-18-19-2-20-15-9-14-11-1	1.619443%
MRMRQ	3-10-16-7-6-17-2-8-13-5-1-4-11-12-14-9-21-15-18-19-20	1.625001%
WRA.	17-21-19-18-1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-20	2.505556%
ORI.	1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20-21	2.0500015%
CON.	IN ONE BATCH	1.8638875%

According to Tables 2-5, comparing with all the other approaches based on different criteria, it is obvious that ITID-AD, which was derived by the evolving linear discriminant with the criterion of maximum AD means in a growing feature space, exhibits the lowest error rate.

7 Conclusion

An evolving linear discriminant for feature-based incremental learning is presented in this paper. This linear discriminant has the same basic concept with FLD, but it can be employed for pattern classification with IAL in a growing feature space. With the criterion of maximum means of AD during the process of feature importing, such a linear discriminant is applicable to search the optimal feature ordering by an evolutionary searching algorithm. Experimental results showed that the approach using the evolving linear discriminant with the criterion of maximum discriminability mean is more feasible to reduce the classification error rate than some other methods. It indicates that the evolving linear discriminant demonstrated in this paper exhibits effective performance with the discriminability's maximum mean criterion and evolutionary searching algorithm. For classification problems with an increasing dimensional feature space to be solved with IAL, the approaches presented in this paper can be regarded as a solution likely to get better performance.

Acknowledgement. This research is supported by National Natural Science Foundation of China under Grant 61070085.

Reference

1. Guan, S. U. and Liu, J. Incremental Ordered Neural Network Training. Journal of Intelligent Systems, vol. 12, no. 3,137-172(2002).
2. Guan, S. U. and Liu, J. Incremental Neural Network Training with an Increasing Input Dimension. Journal of Intelligent Systems, vol. 13, no.1, 43-69 (2004).
3. Wang, T., Guan, S. U., and Liu, F. Ordered Incremental Attribute Learning based on mRMR and Neural Networks. International Journal of Design, Analysis and Tools for Integrated Circuits and Systems, vol. 2, no.2, 86-90 (2011).

4. Guan, S. U. and Li, S. C. Incremental Learning with Respect to New Incoming Input Attributes. Neural Processing Letters, vol. 14, no.3, 241–260 (2001).
5. Guan, S. U. and Zhu, F. M. An Incremental Approach to Genetic Algorithms Based Classification. IEEE Trans. on Systems, Man and Cybernetics Part B, vol. 35, no. 2, 227-239 (2005).
6. Chao, S., Wong F. An incremental decision tree learning methodology regarding attributes in medical data mining. Proceedings of the Eighth Int'l Conference on Machine Learning and Cybernetics, pp.1694-1699,Baoding, (2009).
7. Liu, X. W., Zhang, G. M., Zhan, Y. B., and Zhu, E. An Incremental Feature Learning Algorithm Based on Least Square Support Vector Machine. Lecture Notes in Computer Science, 5059, 330-338 (2008).
8. Bermejo, P., Ossa, L., Gámez, J. A. and Puerta, J. M.. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. Knowledge-Based Systems, vol.25, no.1, 35-44 (2012).
9. Wang, T., Guan, S.U., Liu, F. Feature Discriminability for Pattern Classification Based on Neural Incremental Attribute Learning. In: Wang, Y.L. and Li, T.R. (Eds.)Foundations of Intelligent Systems, The 2011 Int'l Conf. on Intelligent Systems and Knowledge Engineering. pp.275-280, Springer, Heidelberg, (2011).
10. Peng, H., Long, F., and Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, 1226-1238 (2005).