ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

PCA-based algorithm for constructing ensembles of feature ranking filters

Andrey Filchenkov, Vladislav Dolganov and Ivan Smetannikov*

ITMO University - International Laboratory "Computer technology"
Kronverksky Ave. 49 St. Petersburg 197101 Russia

Abstract. Feature filtering algorithms are commonly used in feature selection for high-dimensional datasets due to their simplicity and efficacy. Each of these algorithms has its own strengths and weaknesses. Ensemble of different ranking methods is a way to provide a stable and efficacious ranking algorithm. We propose a PCA-based algorithm for filter ranking algorithms ensemble. We compared this algorithm with four other rank aggregation algorithms on five different datasets used in the NIPS-2003 feature selection challenge. We evaluated the stability of the resulting rankings and the AUC score for four classifiers learnt on resulting feature sets. The proposed method has shown better stability and above-average efficacy.

1 Introduction

Feature selection is one of two general approaches to reduce the number of features in a model by selecting a subset of relevant features. Feature selection algorithms include [2] embedded methods, wrappers and filters. Filters are subdivided [3] into space search methods and ranking methods. There are many fields and applications, such as targeted advertising, social network analysis, and bioinformatics, where it is necessary to classify large amounts of data [1]. Since the number of features in real-world problems can reach hundreds of thousands, ranking filters are useful, because they are known to be the fastest feature selection techniques.

Algorithms ensemble approach is widely used in machine learning beacause usually there is no best algorithm. This approach was applied for feature selection algorithms [4, 5, 6]. The ranking filters ensemble constracting problem can be reduced to the rank aggregation problem, which also arises in various fields, such as social choice theory [7] or DNA microarray meta-analysis [8].

In this paper we propose a novel approach to construst ensemble of ranking filters and a novel feature ranking algorithm implementing this approach. The proposed algorithm is tested on five datasets and compared with other methods based on rank aggregation.

*This work was financially supported by the Government of Russian Federation, Grant 074-U01.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

2 Basic concepts and notation

In the binary classification problem the input data is composed of a dataset $X = \{x_1, \ldots, x_m\}$ with m objects and a target vector $Y = (y_1, \ldots, y_m)$, where each value is equal either to 1 or to 0. We will use the term *attribute* to refer to features of a given dataset. Let $A = \{a_1, \ldots, a_n\}$ denote the attribute set. Each object of X is described by n attributes from A. Let $x_{ij} = a_j(x_i)$ denote the j-th attribute of the i-th object.

A ranking filter f is a feature selection algorithm that is defined by a scoring function $s : A \to \mathbb{R}$ and a cutting rule κ. This rule for a given ranking or a score set returns an attribute subset. Usually κ chooses a fixed number of high-ranked features (so called "top-k") or is threshold-based. Ranking filter algorithm consists of the following steps:

1. Evaluate the scoring function $s(a_j)$ for each attribute.
2. Rank the attributes according to their scores.
3. Select top attributes according to the cutting rule κ.

Let $F = \{f_1, \ldots, f_k\}$ denote the set of ranking filters, each with a scoring function s_i and a cutting rule κ_i. A method for ensemble construction obtains an attribute subset A' by combining filters from F. A common way to achieve this is to aggregate rankings $\{r_i\}_{i=1}^{k}$ into one and apply a cutting criteria to it.

3 First Principal Component Projection Score Algorithm

The problem of filters ensemble construction from the set F can be reduced to the problem of finding s^* — a combination of scoring functions. Since rank aggregation methods do not employ scores, this approach is more general.

Each attribute a_j can be considered as an object, described with k features $(s_1(a_j), \ldots, s_k(a_j))$. Finding a proper mapping from $(s_i(a_j))_{j=1,i=1}^{j=n,i=k}$ to $(s^*(a_j))_{j=1}^{n}$ might be considered as a dimension reduction problem. Applying feature selection techniques in this problem is equivalent to choosing the best given filter. For obtaining an ensemble, feature extraction approaches should be applied. One of the most popular dimension reduction techniques is the Principal Component Analysis (PCA) [9, 10]. Since the desired number of dimensions is one, only the first principal component (FPC) is required, therefore the problem is equivalent to the original Pearson's finding "best-fitting straight line" — a line, on which the sum of all points squared projections is minimal.

The main idea of the algorithm we call First Principal Component Projection Score (FPCPS) is to use FPC coordinates as new scores to obtain a resulting ranking. Redundancy of data is a theoretical requirement for reasonable PCA application. An empirical study has uncovered that application of most pairs of scoring functions results in approximately the same redundancy in each dataset.

It must be noted that the proposed algorithm should be applied after using the cutting rule for all the ranking filters. Therefore, PCA is applied for a dataset with n' objects described with k features, where n' is the number of attributes after cutting and k is the number of ranking filters in the ensemble.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

4 Experiments

4.1 Experiment design

For experiments we use five datasets (subsection 4.2). Each dataset is randomly split 10 times into a training set (60%) and test set (40%). Thus, we conduct 10 experiments. In each experiment, each of the four basic scoring functions (subsection 4.3) is evaluated for each attribute in the training set. Thus, we obtain four basic rankings. For each experiment, choice of the number of selected attributes is based on attribute score variance, we retain from 10% to 25% of original attributes.

Then we use four methods (subsection 4.4) and the FPCPS algorithm to obtain derivative rankings. For FPCPS, we normalize scores within each experiment for each dataset, since it is a common recommendation for PCA application.

We compare methods using two criteria: the stability of obtained ranking and the efficacy of classification algorithms on the resulting feature subset.

For stability measuring we use an adaptation [11] of the Tanimoto distance. For each dataset and each algorithm we evaluate the stability of resulting rankings by comparing the subsets selected in each cycle of cross-validation. The final stability of an algorithm on a given dataset is calculated as an average of 45 paired comparisons on this dataset (for each pair of 10 experiments).

To measure each ranking efficacy we use the four classifiers (subsection 4.5). For each classifier we evaluate the area under the ROC curve (AUC) for every obtained attribute subset. The greater the AUC metric is, the better ranking filter is.

4.2 Datasets

The five datasets are taken from the NIPS 2003 feature selection challenge[1]. They are preprocessed versions of datasets from the UCI Machine Learning Repository[2]. All datasets have a binary target vector. The dataset names are: Arcene (number of objects $m = 200$; number of attributes $n = 10$K), Gisette ($m = 7$K; $n = 5$K), Dexter ($m = 600$; $n = 20$K), Dorothea ($m = 1150$; $n = 100$K), and Madelon ($m = 2600$; $n = 500$).

4.3 Basic scoring functions

We take four different scoring functions, each is based on comparison of the attribute vector with the target vector Y. Two of them are well-known [4]: it is the Spearman correlation coefficient (SP), which is defined as the Pearson correlation coefficient between ranked variables, and the Symmetrical uncertainty (SU), which is normalized information gain measure. We describe the two other scoring functions in a more detailed way.

[1] http://www.nipsfsc.ecs.soton.ac.uk
[2] http://archive.ics.uci.edu/ml/

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

The Value Difference Metric (VDM) [12] is defined as follows:

$$\mathrm{VDM}(A_j, Y) = \frac{1}{2} \sum_i \left| p(A_j = x_{ij} | y_i = 1) - p(A_j = x_{ij} | y_i = 0) \right| .$$

Next, the Fit criterion (FC) [12] is a measure similar to the z-score used in statistics. A binary variable which determines whether a point x belongs to distribution B_0 or distribution B_1 is defined as

$$\mathrm{FCP}(x, B_0, B_1) = \mathrm{argmin}_{l=0,1} \frac{|x - \bar{B}_l|}{\mathrm{var}(B_l)}.$$

The Fit criterion value is the mean of all such values, where B_0 and B_1 are random variables with conditional probability distribution $p(a_j(x_i) = x_{ij} | y_i = 0)$ and $p(a_j(x_i) = x_{ij} | y_i = 1)$ respectively:

$$\mathrm{FC}(A_j, Y) = \frac{1}{m} \sum_{i=1}^{m} \left[\mathrm{FCP}(x_{ij}, B_0, B_1) = y_i \right] ,$$

where $[a = b]$ equals 1 if $a = b$ and 0 otherwise.

4.4 Rank aggregation algorithms

Several well-known rank aggregation methods are used in experiments: the Borda method and two variants of the Markov Chain method [13]. We also use a simple approach we call Best-Go-First (BGF) algorithm.

Suppose we have attributes $\{a_1, \ldots, a_n\}$ and rankings $R = \{r_1, \ldots, r_k\}$ from k different feature ranking methods. Let $L = \{l_{ij}\}$ be a matrix, where $l_{ij} = |\{r | r \in R, r(i) > r(j)\}|$. For the Borda method we obtain a ranking with a new scoring function $s_{\mathrm{B}}(i) = \sum_{j=1}^{m} l_{ij}$. For the Markov chain methods (MC) $s_{\mathrm{MC}}(i) = 1 - \sum_{i \neq j} P(i \to j)$. In MC1 $P(i \to j) = 1/m$ if $r_t(i) > r_t(j)$ in at least one ranking, and 0 otherwise. In MC2 $P(i \to j) = 1/m$ if $r_t(i) > r_t(j)$ for majority of rankings, and 0 otherwise. In the BGF method a new ranking r^* is obtained by taking iteratively at each step t all attributes $\{a_{[h,t]}\}_{h=1}^{h=k}$ such that $r_h(a_{[h,t]}) = t \ \forall h = 1, \ldots, k$ and $a_{[h,t]} \notin r^*$.

4.5 Classifiers

We take four commonly used classifiers: Naive Bayes, k-Nearest Neighbors (IBk), Random Forest and Support Vector Machine (SMO). The implementation of these algorithms was taken from Weka. All the algorithms are learnt with default parameters except IBk. We set CrossValidate parameter for IBk to select optimal value of k with cross validation.

5 Results

As shown in Table 1, for almost all datasets the value of FPCPS algorithm stability surpasses all the values of aggregation algorithms except MC2.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

	Dorothea	Gisette	Dexter	Arcene	Madelon
Borda	0.410	0.815	0.335	0.600	0.256
MC1	0.380	0.813	0.445	0.513	0.241
MC2	0.435	**0.842**	0.469	**0.613**	0.263
BGF	0.447	0.467	**0.471**	0.556	0.260
FPCPS	**0.471**	0.755	0.436	0.564	**0.288**

Table 1: Stability of aggregation algorithms.

The means of AUC metric value for the Madelon dataset are shown in Table 2. Used classifiers are presented in the first column. The first row contains used feature ranking methods. Other cells of the table contain average AUC measure values for corresponding feature ranking methods and classifiers. AUC tables for other datasets can be found in the repository[3].

	SP	SU	VDM	FC	Borda	MC1	MC2	BGF	FPCPS
KNN	0.599	0.612	0.547	0.569	0.594	0.589	0.613	**0.629**	0.627
NB	0.639	**0.643**	0.613	0.636	0.639	0.640	0.638	0.641	0.640
RF	0.751	0.769	0.639	0.698	0.745	0.721	0.765	**0.779**	0.774
SVM	0.580	**0.612**	0.609	0.589	0.586	0.593	0.581	0.596	0.596

Table 2: AUC scores on Madelon dataset.

We use paired Wilcoxon signed-rank test to check whether two feature selection methods are statistically different on the given dataset and using a given classifier. Two methods are considered statistically distinguishable if the p-value is less than 0.05. For example, FPCPS and BGF are not statistically different for the SVM classifier trained on the Madelon dataset. The tables of p-values for classifiers and datasets can be found in the repository[4].

We estimate nine feature selection methods with different ranking algorithms on five datasets. Experiments show that the proposed method is among the top three ones with highest AUC scores or it is statistically indistinguishable from such methods on all datasets except the Gisette dataset. On the Gisette dataset, FPCPS has the fourth result in average.

6 Conclusion and Future Work

We have proposed the FPCPS algorithm for ensemble construction. It builds a scoring function as an ensemble of scoring functions using PCA. Applied to four datasets, the proposed algorithm showed encouraging stability and appeared to be among top-3 most efficacious algorithms for all the classifiers.

If any two coefficients in the FPC equation have different signs, projections of all high-scored attributes will be situated between lower-scored ones, therefore final ranks of high-scored attributes will not be the highest. This problem may occur when many lower-scored but highly dispersed attributes have more impact on the first principal component equation than others. Experiments

[3]http://genome.ifmo.ru/files/papers_files/ESANN2015/AUCs.pdf
[4]http://genome.ifmo.ru/files/papers_files/ESANN2015/p-values.pdf

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

show that this situation never arises after applying the cutting rule. We hold special experiments with the cutting rule applied after merging all scoring functions, therefore all the attributes are used to find FPC. In the Dorothea dataset, which has the largest number of attributes, the FPCPS algorithm shows extremely poor results, while in other datasets the results are slightly worse than in the experiments with preliminary application of the cutting rule.

The proposed approach has shown a potential of further improvement. We suggest that there are three possible directions of future work. First, to solve the reformulated problem with attributes weighted according to their score. Weighted PCA [14] can also be use to solve this problem. The problem of finding the attribute weight function is similar to the problem of finding a resulting scoring function with a difference that these functions are applied for different purposes. Second, to learn coefficients in the linear equation of resulting scoring function. Finally, to apply non-linear approaches for combining scoring functions.

References

[1] I. A. Gheyas, L. S. Smith, Feature subset selection in large dimensionality domains, *Pattern recognition*, 43:5–13, 2010.

[2] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[3] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, A. Nowe, A survey on filter techniques for feature selection in gene expression microarray analysis *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9:1106–1119, 2012.

[4] Y. Saeys, T. Abeel, Y. Van de Peer, Robust feature selection using ensemble feature selection techniques, *ML and KDD*, 313–325, 2008.

[5] A. Tsymbal, M. Pechenizkiy, P. Cunningham, Diversity in search strategies for ensemble feature selection, *Information fusion*, 6:83–98, 2005.

[6] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, An ensemble of filters and classifiers for microarray data classification, *Pattern Recognition*, 45:531–539, 2012.

[7] Y. Chevaleyre, U. Endriss, J. Lang, N. Maudet, A short introduction to computational social choice, *SOFSEM 2007: Theory and Practice of Computer Science*, Lecture Notes in Computer Science, pages 51–69, Springer Berlin Heidelberg, 2007.

[8] A. L. Boulesteix, M. Slawski, Stability and aggregation of ranked gene lists. *Briefings in bioinformatics*, 10:556–568, 2009.

[9] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, R. Tibshirani, *The elements of statistical learning (2nd ed.)*, Springer, New York, 2009.

[10] I. Jolliffe, *Principal component analysis*, John Wiley & Sons, Ltd, 2005.

[11] A. Kalousis, J. Prados, M. Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, *J. of Knowledge and Information Systems*, 12:95–116, 2007.

[12] B. Auffarth, M. López, J. Cerquides, Comparison of redundancy and relevance measures for feature selection in tissue classification of CT images, *Advances in Data Mining. Applications and Theoretical Aspects*, 248–262, 2010.

[13] S. Lin, Rank aggregation methods, *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:555–570, 2010.

[14] D. Skocaj, A. Leonardis, Weighted incremental subspace learning, *The proceeding of European workshop on Cognitive Vision*, Zürich, Switzerland, 19–20, 2002.