PCA dimensionality reduction for categorical data

Aleksander Denisiuk[0000−0002−7501−7048]

University of Warmia and Mazury in Olsztyn
ul. Słoneczna 54, 10-710 Olsztyn, Poland
denisiuk@matman.uwm.edu.pl

Abstract. The purpose of the article is to develop a new dimensionality reduction algorithm for categorical data. We give a new geometric formulation of the PCA dimensionality reduction method for numerical data that can be effectively transferred to the case of categorical data with the Hamming metric.

Keywords: PCA · Hamming metric · weighted Hamming metric · categorical data · dimensionality reduction · classification.

1 Introduction

One of the objectives of principal component analysis (PCA) is to reduce the dimension of the data space while retaining as much information as possible. The standard algorithm (see, for instance [12]) consists of calculating the eigenvectors of the covariant (correlation) matrix and using it as a new basis in the space of data. Coordinates of data vectors in this new basis are the principle components. The major principle component corresponds to the eigenvector with the largest eigenvalue, the minor component—to the eigenvector with the smallest eigenvalue. The dimensionality reduction involves discarding minor components.

However this algorithm is not applicable in case of categorical data, where the structure of linear space is not available.

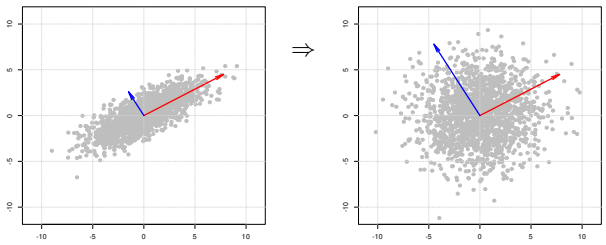

Fig. 1. Scaling that minimizes total relative inner-class squared distance

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_22

In this article we propose a new interpretation of the PCA dimensionality reduction and transfer it to a set of categorical data. Namely, consider an affine transform of the data space that minimizes the total relative squared inner-class distance (figure 1). It turns out that the major principle component will be scaled with the minimal multiplier, while minor component with the maximal one.

The problem of dimensionality reduction can be brought to finding a scaling that minimizes the relative total squared distance. The direction with the most scaling multiplier corresponds to the minor component that can be dropped with a minimal information lost.

Such interpretation can be transferred to a space of categorical data with the weighted Hamming metric.

To prove the concept we perform numeric experiments on three datasets. Experiments show that discarding the features in order according to our method always results in the loss of a minimum amount of information. Discarding the features in reverse order results in maximal information loss.

The rest of the paper is organized as follows. In section 2 we shortly recall basic related works paying main attention to recent works concerning non-numeric data. Section 3 contains our new interpretation of PCA dimensionality reduction for numeric data. This interpretation in transferred to categorical data in the section 4. To verify our concept we performed numerical experiments that are described in the section 5. Finally, we give some concluding remarks in the section 6.

2 Related Works

Developed for numerical continuous data in the pioneering works of Pearson [18] and Hotelling [11], PCA has found many applications in many fields of data analysis. The method and recent developments for continuous numerical data are described in book [12] and review [13].

We focus on extensions of the PCA to discrete data. The most known is the correspondence analysis (see, for instance [12, section 5]) which deals with the principal components of the normalized contingency matrix.

In articles [10, 7] the PCA was applied to a binary data. In case of ordinal data the authors of [14] suggested a variant of PCA based on Spearman's and Kendall's rank correlation coefficients. In our algorithm, the data features re not ordered and can have arbitrary cardinalities.

Let us also mention some recent papers that develop PCA for discrete data with additional complex structure as intervals or histograms [16, 2, 4]. In our parer we do not assume any additional structure defined on the data.

The PCA is often formulated as an optimization problem: maximizing dispersion of data projection, optimal approximation of the data with a linear manifold, finding projection that maximize the total inner-class squared distance. To the best of our knowledge the optimization problem suggested in this paper was not directly considered before. A characteristic feature that distinguishes our algorithm from other approaches: first we define a minor feature, and the other algorithms start by determining the most important feature.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_22

The weighted Hamming metric itself was recently used for unsupervised [9] and supervised [8] metric learning for numeric-categorical data. Application of the weighted Hamming metric to the problem of dimensionality reduction of categorical data seems to be new.

3 Reformulation of PCA dimensionality reduction for numerical data

In this section we will show that minimizing of the total relative squared inner-class distance allows us to determine the minor principle component.

Assume that each data instance x has n numerical features, i.e. $x \in \mathbb{R}^n$ with the standard Euclidean distance. The distance is invariant with respect to translations and rotations. So, let us make two following assumptions:

1. The features are uncorrelated,
2. The data has a multivariate normal distribution centered at the origin.

That means that distribution function is as follows:

$$f(x) = \frac{\exp\left(-\frac{1}{2}x^T \Sigma^{-1} x\right)}{\sqrt{(2\pi)^n \det \Sigma}},$$

where Σ is the correlation matrix, which in case of uncorrelated data is diagonal $\Sigma = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $\Sigma^{-1} = \operatorname{diag}(\lambda_1^{-1}, \ldots, \lambda_n^{-1})$, $\det \Sigma = \lambda_1 \cdots \lambda_n$, and $\lambda_i > 0$ for $i = 1, \ldots, n$. The total inner-class squared distance is

$$H = \int_{\mathbb{R}^n \times \mathbb{R}^n} \operatorname{dist}^2(x, y) f(x) f(y)\, dx\, dy,$$

where $dx = dx_1 \ldots dx_n$, $dy = dy_1 \ldots dy_n$.

We are to find a scaling $(x_1, \ldots, x_n) \mapsto (w_1 x_1, \ldots, w_n x_n)$, where $w_i \geq 0$ for $i = 1, \ldots, n$, that minimizes the total inner-class squared distance:

$$H(w) = \sum_{i=1}^{n} w_i^2 \int_{\mathbb{R}^n \times \mathbb{R}^n} (x_i - y_i)^2 f(x) f(y)\, dx\, dy. \tag{1}$$

Since the function $H(w)$ is homogenious we add a constraint on the weight w:

$$\sum\nolimits_{i=1}^{n} w_i = 1. \tag{2}$$

The restriction of $H(w)$ to (2) is the *total relative inner-class squared distance.*

By the standard calculation the coefficient at w_i^2 in (1) equals $z_i = 2\lambda_i$. So, the minimization problem is as follows:

$$\begin{cases} \sum_{j=1}^{n} w_j^2 \lambda_j \to \min, \\ \sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0 \text{ for } j = 1, \ldots, n. \end{cases}$$

One can solve it with the method of Lagrange multipliers:

$$w_i = \left(\sum\nolimits_{j=1}^{n} \lambda_j^{-1}\right)^{-1} / \lambda_i, \quad i = 1, \ldots, n.$$

Specifically, the minor component (of minimal λ_i) has the maximal multiplier.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_22

4 Dimensionality reduction for categorical data

In this section we transfer considerations from the last section to the case of categorical data.

Assume that the dataset $\mathbf{X}$ of M instances is given. Let each instance $x \in \mathbf{X}$ has n categorical features of finite cardinalities $a_1, \dots, a_n$ respectively, $x = (x_1, \dots, x_n)$. We use the standard Hamming metric as a distance on $\mathbf{X}$:

$$\mathrm{dist}_h(x, y) = \sum\nolimits_{i=1}^{n} \mathrm{diff}(x_i, y_i),$$

where $x, y \in \mathbf{X}$, and $\mathrm{diff}(\alpha, \beta) = \begin{cases} 1, & \text{if } \alpha \neq \beta, \\ 0, & \text{if } \alpha = \beta. \end{cases}$

Let us also make an assumption that the dataset is divided into c classes, $\mathbf{X} = C_1 \cup \dots \cup C_c$. The total inner-class distance is

$$G = \frac{1}{M^2} \sum_{k=1}^{c} \sum_{x,y \in C_k} \mathrm{dist}_h(x, y).$$

Do define a "scaling", let us introduce the weights vector $u = (u_1, \dots, u_n) \in \mathbb{R}^n$, where $u_i \geq 0$ for $i = 1, \dots, n$ and the weighted Hamming distance:

$$\mathrm{dist}_{h,u}(x, y) = \sum\nolimits_{i=1}^{n} u_i \, \mathrm{diff}(x_i, y_i).$$

Consider minimization problem for the function

$$G(u) = \frac{1}{M^2} \sum_{k=1}^{c} \sum_{x,y \in C_k} \mathrm{dist}_{h,u}(x, y) = \sum_{i=1}^{n} u_i \left(\frac{1}{M^2} \sum_{k=1}^{c} \sum_{x,y \in C_k} \mathrm{diff}(x_i, y_i) \right) \quad (3)$$

with the following constraint on weights u:

$$\sum\nolimits_{i=1}^{n} u_i = 1, \quad (4)$$

which is exactly the same as (2), and call the restriction of $G(u)$ to the hyperplane (4) the *total relative inner-class distance.*

Denoting coefficient at u_i in (3) by s_i, we obtain the following minimization problem:

$$\begin{cases} \sum\limits_{i=1}^{n} u_i s_i \to \min, \\ \sum\limits_{i=1}^{n} u_i = 1, \quad u_j \geq 0 \text{ for } i = 1, \dots, n, \end{cases} \quad (5)$$

which is known as the linear programming problem. The objective function reaches its optimal value at one of the vertices of the polytope defined by the constraint. Hence the optimal vector has the form $u_{\mathrm{opt}} = (0, \dots, 0, 1, 0, \dots 0)$, i.e. all the coordinates but one are zeroes. The feature k that corresponds to coordinate $u_{\mathrm{opt},k} = 1$ is called *the minor component* and can be discarded first in dimensionality reduction.

Repeating this procedure after the feature k reduction, one determines the next minor feature, and so on. We summarize the method in the algorithm 4.1.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_22

Algorithm 4.1 Reduction of m dimensions

Require: the dimension of dataset $\mathbf{X}$ is n, $n > m$
Ensure: the dimension of dataset $\mathbf{X}$ is $n - m$
  $s \leftarrow 0$
  while $s < m$ **do**
    solve the optimization problem (5) and discard the minor component
    $s \leftarrow s + 1$
  end while

5 Numerical Experiments

To illustrate the concept a few R scripts have been created. The code is available as a project on Gitlab at `https://gitlab.com/adenisiuk/pca`.

The purpose of our test is to show that the algorithm allows to reduce dimensionality while retaining as much information as possible. To do this we consider the classification problem. We classified data with complete set of features and then we discarded features one by one in different orders. First, according to the proposed algorithm: starting from the most minor feature, this order is referred as *pca* order (red line on the figures). Second order is reverse to the pca order: starting from most major features. We call this order *acp* order (blue line on the figures). We also performed two tests with random order of features discarding, the *sample* order (green lines on the figures).

Implementations of three classifiers: random forest, SVM and XGBoost in R were used in experiments: [15, 17, 5]. For the random forest classifier an average result for 100 tests is presented.

We use the F_1Score as a measure of classification accuracy. Some of datasets have more than two classes. In all the tests we define F_1Score as

$$\mathrm{F}_1\mathrm{Score} = \frac{2 \cdot \mathrm{total\_presicion} \cdot \mathrm{total\_recall}}{\mathrm{total\_presicion} + \mathrm{total\_recall}},$$

where total_presicion and total_recall are the sums of respectively presicion and recall for all the classes.

We considered the following three datasets: the Car Evaluation [3], the Congressional Voting Records [6] and the Tic-Tac-Toe Endgame [1]. All the tested datasets were split into train (80%) and test (20%) parts.

One can see that in all the tests the *pca* order always gives the minimal information loss while the *acp* order gives the maximal lost.

5.1 Car Evaluation dataset

The dataset contains 1728 instances. Each instance has 6 features of cardinalities respectively $4, 4, 4, 3, 3, 3$. The dataset is split into 4 classes [3].

The results of classification are presented at the figure 2. The *pca* order for this dataset is 3, 2, 1, 5, 4, 6. Orders sampled in tests were as follows: 4, 1, 3, 2,

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_22

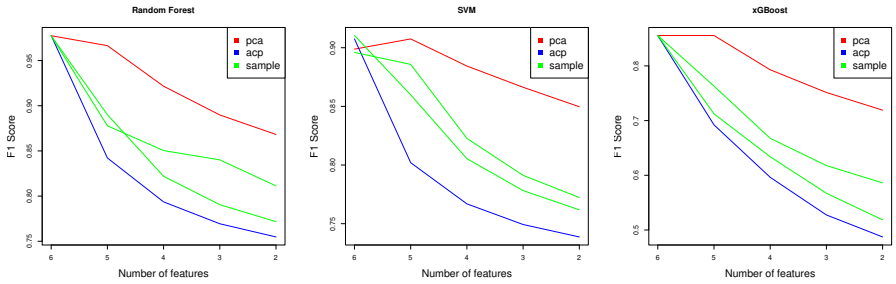

Fig. 2. Classification accuracy for the Car Evaluation dataset

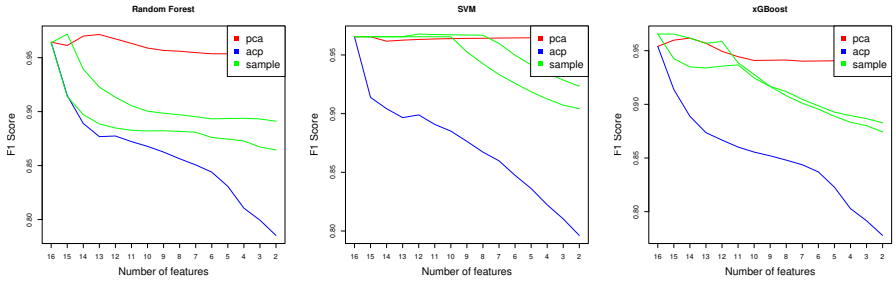

Fig. 3. Classification accuracy for the Congressional Voting Records dataset

6, 5 and 1, 5, 3, 6, 4, 2 for the random forest classifier, 1, 6, 4, 3, 5, 2 and 5, 4, 3, 1, 6, 2 for SVM, 2, 4, 1, 3, 5, 6 and 4, 2, 6, 3, 5, 1 for XGBoost.

One can observe the for all classifiers the *pca* order has minimal and the *acp* order has maximal accuracy loss when discarding features.

5.2 Congressional Voting Records dataset

The dataset contains 435 instances. Each instance has 16 features of cardinalities 3. According to the data description we interpreted values "?" as the third option in voting. The dataset is split into 2 classes [6].

The results of classification are presented at the figure 3. The *pca* order for this dataset is 2, 10, 16, 11, 1, 15, 13, 6, 14, 9, 7, 12, 5, 8, 3, 4. Orders sampled in tests were as follows: 4, 16, 6, 10, 7, 1, 5, 14, 2, 3, 12, 13, 11, 9, 8, 15 and 10, 4, 8, 6, 16, 5, 2, 7, 15, 14, 9, 1, 12, 13, 11, 3 for the random forest classifier, 1, 11, 7, 9, 14, 2, 4, 13, 8, 10, 16, 6, 15, 5, 3, 12 and 9, 10, 5, 8, 12, 16, 2, 14, 4, 11, 15, 13, 6, 7, 3, 1 for SVM, 12, 15, 7, 9, 4, 11, 1, 8, 14, 3, 16, 10, 6, 13, 2, 5 and 11, 16, 5, 7, 13, 4, 9, 14, 2, 6, 3, 8, 10, 15, 12, 1 for XGBoost.

Again, the *pca* order has minimal and the *acp* order has maximal accuracy loss when discarding features. Note also very stable classification for the *pca* order and drastic drop in accuracy after discarding the component 4, which is

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_22

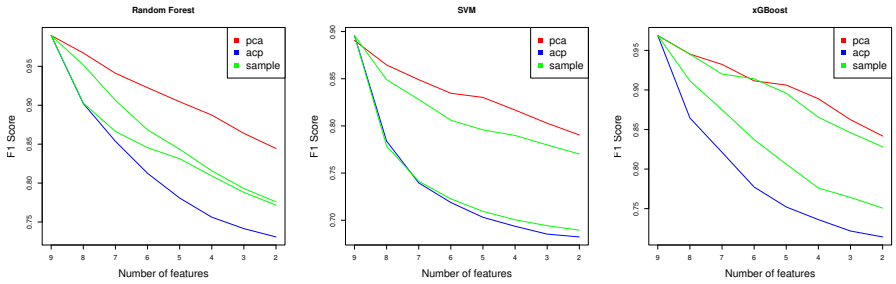

Fig. 4. Classification accuracy for the Tic-Tac-Toe Endgame

most significant according to our algorithm. This is especially noticeable at the *sample* orders for random forest classifier: the 4th feature was the first and the second one.

5.3 Tic-Tac-Toe Endgame dataset

The dataset contains 958 instances. Each instance has 16 features, each one is of cardinality 3. The dataset is split into 2 classes [1].

The results of classification are presented at the figure 4. The *pca* order for this dataset is 2, 4, 8, 6, 9, 3, 7, 1, 5. Orders sampled in tests were as follows: 5, 6, 2, 3, 4, 8, 1, 9, 7 and 4, 9, 3, 6, 5, 1, 8, 7, 2 for the random forest classifier, 5, 4, 7, 3, 9, 2, 8, 1, 6 and 7, 8, 3, 6, 4, 9, 1, 2, 5 for SVM, 2, 8, 6, 3, 1, 7, 9, 5, 4 and 3, 4, 7, 5, 2, 1, 9, 6, 8 for XGBoost.

As in two previous experiments, the *pca* order has minimal and the *acp* order has maximal accuracy loss when discarding features. The accuracy graphs for this dataset is more monotonic than for the Congressional Voting Records. This is probably related to greater uncorrelatedness of the features. As in the previous example, let us notice drop in accuracy after discarding the feature 5: see two *sample* orders for the random forest and the SVM classifiers where this feature is the first one.

6 Conclusion and Future Work

In this article we propose a new geometric interpretation of the PCA dimensionality reduction algorithm and transfer it to categorical data. The algorithm involves determining and discarding minor features.

Numerical experiments confirm that the method allows to discard features with minimal lost of information, at least for the classification problem.

So, we can suggest this method of dimensionality reduction for categorical data analysis.

Hovewer, unlike the classical PCA, our method does not provide any quantitative characterization of the amount of information discarded. Developing such an interpretation is one of the future tasks.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_22

Another direction of the future work is extension of the method to data that have both numerical and categorical features.

References

1. Aha, D.: Tic-Tac-Toe Endgame. UCI Machine Learning Repository (1991)
2. Bock, H. H., Diday, E.: Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer Science & Business Media (2012)
3. Bohanec, M.: Car Evaluation. UCI Machine Learning Repository (1997),
4. Brito, P.: Symbolic data analysis: another look at the interaction of data mining and statistics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **4**(4), 281–295 (2014)
5. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J.: XGBoost: Extreme Gradient Boosting (2023),
6. Congressional Voting Records. UCI Machine Learning Repository (1987),
7. Cox, D.R.: The analysis of multivariate binary data. Applied statistics pp. 113–120 (1972)
8. Denisiuk, A.: Weighted hamming metric and knn classification of nominal-continuous data. In: Mikyška, J., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V. V., Dongarra, J. J., Sloot, P. M. (eds.) Computational Science – ICCS 2023. pp. 306–313. Springer Nature Switzerland, Cham (2023)
9. Denisiuk, A., Grabowski, M.: Embedding of the hamming space into a sphere with weighted quadrance metric and c-means clustering of nominal-continuous data. Intelligent Data Analysis **22**(6), 1297001314 (2018).
10. Gower, J. C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**(3-4), 325–338 (12 1966).
11. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of educational psychology **24**(6), 417 (1933)
12. Jolliffe, I. T.: Principal component analysis. Springer (2002)
13. Jolliffe, I. T., Cadima, J.: Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **374** (2016),
14. Korhonen, P., Siljamäki, A.: Ordinal principal component analysis theory and an application. Computational Statistics & Data Analysis **26**(4), 411–424 (1998).
15. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002)
16. Makosso-Kallyth, S.: Principal axes analysis of symbolic histogram variables. The ASA Data Science Journal **9**(3), 188–200 (2016).
17. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2022), r package version 1.7-12
18. Pearson, K.: On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science **2**(11), 559–572 (1901)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_22