Rafał ŁOPATKA, Dominik OLSZEWSKI

Warsaw University of Technology

Employing Probabilistic Dissimilarity for Feature Discovery in a Game of Chess

Abstract. *We present the feature discovery technique based on the use of the probabilistic dissimilarity, i.e., a measure of dissimilarity between two probability distributions. The solutions in the field of feature discovery, generally, fall into feature extraction and feature selection methods. Both of these groups form the feature subset on the basis of the initial feature set. Also, both of the groups use the numeric representations of features, what often can be misleading, since the different physical meaning of different features can be lost, when they are all treated only as numbers. The approach we propose does not require the initial feature set. Moreover, it does not require the numeric representation of the features. Instead, we propose using only one numeric, decimal quantity allowing for effective feature discovery. We demonstrate that taking advantage of the probabilistic dissimilarity during the feature retrieval phase can benefit by discovering relevant features. We show the way to create a probabilistic model of the analyzed data set, required for the use of the proposed technique. Finally, we report the experimental results of application of the feature discovery method introduced in this paper to the game of chess.*

Streszczenie. *Przedstawiamy technikę odkrywania cech wykorzystującą pseudoodległość probabilistyczną, będącą miarą podobieństwa pomiędzy dwoma rozkładami prawdopodobieństwa. Rozwiązania zaproponowane w dziedzinie odkrywania cech mogą być w ogólności podzielone na metody ekstrakcji i selekcji cech. Obie te grupy metod formują podzbiór cech na podstawie początkowego zbioru cech. Obie te grupy wykorzystują również reprezentacje liczbowe cech, co często może być mylące, gdyż różne znaczenie fizyczne różnych cech może zostać utracone, kiedy wszystkie cechy traktowane są jedynie jako liczby. Proponowane podejście nie wymaga początkowego zbioru cech. Co więcej, nie wymaga ono reprezentacji liczbowej cech. W zamian proponjemy wykorzystanie tylko jednej, dziesiętnej wielkości liczbowej, pozwalającej na skuteczne odkrywanie cech. Demonstrujemy, że wykorzystanie pseudoodległości probabilistycznej pozwala odkryć istotne cechy. Przedstawiamy także sposób budowy modelu probabilistycznego analizowanych danych, wymaganego do zastosowania proponowanej techniki. W części pracy poświęconej eksperymentom, przedstawiamy wyniki zastosowania proponowanej metody odkrywania cech w dziedzinie gry w szachy.* ***(Wykorzystanie pseudoodległości probabilistycznej w odkrywaniu cech w grze w szachy)***

Keywords: feature discovery, feature extraction, feature selection, probabilistic dissimilarity, Hellinger distance, game of chess
Słowa kluczowe: odkrywanie cech, ekstrakcja cech, selekcja cech, pseudoodległość probabilistyczna, odległość Hellingera, gra w szachy

Introduction

The probabilistic dissimilarity is a quantity, widely used in probability theory and statistics as a measure of dissimilarity between two probability distributions. We focus on its usage in feature discovery problems. Feature discovery is an important data pre-processing stage having a strong impact on the subsequent analysis, like classification, clustering, or regression, to name a few. It aims to form possibly smallest set of most relevant, discriminative, and informative features, being the most useful representation of the objects in the analyzed data set. Finding different features results in different quality of the subsequent data processing. The choice of the most appropriate features for a particular task is a well-known problem in multivariate analysis. In the case of more complex issues, it is sometimes necessary to consult an expert in a considered domain. The existing approaches to feature discovery cannot assure obtaining of the most desirable feature subset. They choose certain features without any guarantee that they are most useful in the context of a particular problem. Preliminary data analysis with application of the statistical means, introduced in this paper, can essentially help to retrieve features properly characterizing the objects in the considered data set.

Related Work and Our Method

Many solutions have been developed in the field of feature discovery. Generally, they may be divided into two groups: feature extraction [1–4] and feature selection [5–8] methods. Although the feature extraction and feature selection problems have been widely studied, they still remain a challenge, since there exist no optimal method for determining the most relevant features. The feature extraction methods aim to extract features by projecting the original high-dimensional data into a lower-dimensional space through algebraic transformations. The feature selection methods aim to find the representative features from the set of all features according to some criteria, for example the features are ranked according to their predictive power [9]. In other words, feature extraction is considered a process to generate a new and smaller feature set by linearly or nonlinearly combining the original features, while feature selection is an approach to selecting relevant feature subset from the original feature set. Although feature selection preserves the original physical meaning of the selected features, it is considered as more computationally complex, less flexible, and less effective than feature extraction approach [9, 10]. Both, the feature extraction and feature selection approaches belong to the optimization problems class. The difference between them derives from different forms of objective functions. The feature extraction algorithms seek for the solution in a continuous space, while the feature selection algorithms aim to find the solution in a discrete space [9]. The feature extraction and the feature selection tasks are often framed as the dimensionality reduction problems [9–12]. Both, the feature extraction and feature selection methods assume the determination of most wanted features on the basis of the initial feature set, what can be regarded as an inhibiting constraint. They form a feature set, which is the subset of the initial feature set. This requirement is a constraint which does not concern the approach proposed in this paper. Our method works in the opposite way: it forms a feature set on the basis of one, decimal feature, which should be a random variable, and which we will call the quantity associated with the set (see Definition 3). Therefore, the introduced solution allows to determine features without any initial feature set, just on the basis of the original data set and a random quantity, after specific analysis with use of the probabilistic dissimilarity. We have decided for naming the proposed solution as a feature discovery problem to emphasize the relevance of the fact that no initial feature set is required.

On the other hand, the authors of [13] avoid the feature discovery process, instead, using the discrete probability distributions, built according to the procedure introduced in their work.

The feature extraction and feature selection approaches, both, utilize the numeric representation of the features. The comparison of the features on the basis of their numeric rep-

resentation can often lead to incorrect results - the number associated with certain feature should be used, essentially, only referred to this feature. Two popular feature extraction algorithms: Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), and the Class Separability feature selection criterion, for example, are based on the utilization of the covariance-matrices (PCA), or scatter-matrices (LDA, Class Separability), which are created on the basis of feature-vectors. This results in applying the same algebraic transformations to all features, regardless they represent different properties, and, most likely are expressed in different units. The problem become especially noticeable in the case of binary features (the feature describes existence or non-existence of certain property). The numeric (binary) representation of such feature should not be algebraically compared to other features expressed in the decimal numbers for example. This problem is a well-known issue by the wide range of researchers and engineers, and the fact, the proposed solution use only one numeric, always decimal quantity, instead of numeric representations of all features, thus, avoiding associated with it difficulties, may be recognized as a significant advantage.

The Probabilistic Dissimilarity

In the probability theory and statistics, a several probabilistic dissimilarities are proposed. A survey of the frequently used ones can be found, for example, in [14–16]. Some of them are metrics (satisfy all metrics requirements), and some are not, but still present useful properties. Our method does not impose any requirements to the dissimilarity, in other words, an application of any probabilistic dissimilarity is possible. However, in our experiments, we had to choose the specific one, and we have decided for the use of the Hellinger distance. This choice was motivated with the useful and convenient properties of this quantity. Therefore, throughout this paper, we will refer to the Hellinger distance as the probabilistic dissimilarity. In order to define this dissimilarity and its properties, we will use the following notation.

Let $\mathbb{P}$ and $\mathbb{Q}$ denote two probability measures on a measurable space Ω with σ-algebra $\mathcal{F}$. Let λ be a measure on $(\Omega, \mathcal{F})$ such that $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous with respect to λ, with corresponding density functions p and q (for example, λ can be taken to be ($\mathbb{P}$ + $\mathbb{Q}$)/2 or can be the Lebesgue measure).

Definition 1 ([14, 16]) *The Hellinger distance between $\mathbb{P}$ and $\mathbb{Q}$ on a continuous measurable space $(\Omega, \mathcal{F})$ is defined as*

$$\mathrm{H}(\mathbb{P},\mathbb{Q}) := \left[\frac{1}{2}\int\left(\sqrt{\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\lambda}}-\sqrt{\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\lambda}}\right)^2\mathrm{d}\lambda\right]^{1/2} = \tag{1}$$

$$= \frac{1}{2}\int\left(\sqrt{p}-\sqrt{q}\right)^2\mathrm{d}\lambda^{1/2}$$

The square roots of densities $\sqrt{p}$ and $\sqrt{q}$ belong to the Hilbert space of square integrable functions L^2 [15]. This definition does not depend on the choice of the measure λ [15, 16]. Hellinger distance can be used to estimate the distances between two probability measures independent of the parameters.

For a countable space Ω, measures $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ are N-tuples $(p_1, p_2, \ldots, p_N)$ and $(q_1, q_2, \ldots, q_N)$, respectively, satisfying following conditions: $p_i \geq 0$, $q_i \geq 0$, $\sum_i p_i = 1$ and $\sum_i q_i = 1$.

Definition 2 ([16, 17]) *The Hellinger distance between measures $\mathbb{P}$ and $\mathbb{Q}$ on a discrete measurable space $(\Omega, \mathcal{F})$ is defined as*

$$\mathrm{H}(\mathbb{P},\mathbb{Q}) := \left[\frac{1}{2}\sum_{i=1}^{N}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2\right]^{1/2} \tag{2}$$

In some papers [16, 17] the factor of $\frac{1}{2}$ in Definitions 1 and 2 is omitted. We consider definition containing this factor, as it normalizes the range of values taken by the distance. Some sources [18, 19] define the Hellinger distance as the square of H. Defined by formulae (1) and (2) Hellinger distance is a metric, while H^2 is not a metric, since it does not satisfy the triangle inequality.

Hellinger Distance's Properties

The Hellinger distance has the following properties:

1. It takes values from the interval [0, 1] which is convenient, since it can be interpreted as the probability measure (if the factor $\frac{1}{2}$ is removed from the Definitions 1 and 2 values taken by the distance belong to the interval [0, $\sqrt{2}$]) [15, 16].
2. $\mathrm{H}(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\sqrt{p} = \sqrt{q}$, that is, when $\mathbb{P} = \mathbb{Q}$ [15].
3. $\mathrm{H}(\mathbb{P},\mathbb{Q}) = 1$ if and only if $pq = 0$ which is the condition for disjoint $\mathbb{P}$ and $\mathbb{Q}$ [15].
4. It is symmetric, that means $\mathrm{H}(\mathbb{P},\mathbb{Q}) = \mathrm{H}(\mathbb{Q},\mathbb{P})$.
5. It satisfies the triangle inequality, which means that $\mathrm{H}(\mathbb{P},\mathbb{Q}) \leq \mathrm{H}(\mathbb{P},\mathbb{R}) + \mathrm{H}(\mathbb{R},\mathbb{Q})$ for any $\mathbb{P},\mathbb{Q},\mathbb{R}$.
6. For product measures $\mathbb{P} = \mathbb{P}_1 \times \ldots \times \mathbb{P}_n$, $\mathbb{Q} = \mathbb{Q}_1 \times \ldots \times \mathbb{Q}_n$ on a product space $\Omega_1 \times \ldots \times \Omega_n$ [15, 16]

$$\mathrm{H}^2(\mathbb{P},\mathbb{Q}) = 2 - 2\prod_{i=1}^{n}\left(1-\frac{1}{2}\mathrm{H}^2(\mathbb{P}_i,\mathbb{Q}_i)\right) \tag{3}$$

The consequence of (3) is as follows [15]

$$\mathrm{H}^2(\mathbb{P},\mathbb{Q}) \leq \sum_{i=1}^{n}\mathrm{H}^2(\mathbb{P}_i,\mathbb{Q}_i) \tag{4}$$

Factorization presented in (3) is the main reason for great usefulness of the Hellinger distance in problems involving product measures [15].

Building of the Probabilistic Model

The concept of feature discovery process improvement introduced in this paper assumes the application of the Hellinger distance as a probabilistic tool increasing the accuracy of choosing the most appropriate features for the specific issue. This requires building of a statistical representation of the considered data set.

The probabilistic model that we utilized concerned discrete probability distributions that can be associated with the analyzed data set.

Definition 3 *The quantity associated with the set is defined as the quantity, which can be determined for each object in this set. This quantity is assumed to be a random variable.*

For a purpose of building mentioned distributions one need to choose the quantity associated with the analyzed data set. Values of this quantity for each of the objects in the analyzed set are their realizations. Each of the distributions will be constructed on the basis of these realizations.

Definition 4 *The general discrete probability distribution is defined as the discrete probability distribution obtained by choosing randomly a fixed number of objects from the set and building a probability distribution based on the values of the quantity associated with the set for each of the chosen objects.*

The target model will consist of the set of three discrete probability distributions, including the general discrete probability distribution. The other two distributions will be created according to the Procedure 1 in the next section. The selection of the quantity associated with the set, necessary for building of a probabilistic model, strongly impacts on the subsequent stages of our method, and, consequently, on the effectiveness of the entire feature discovery approach proposed in this paper. Therefore, the quantity should be possibly most relevant, discriminative and informative feature of the objects in the analyzed data set. We do not provide an exact, principled way to find it, however, we claim that the advantage of our method is that it leads to discovery of the feature set on the basis of only one quantity. This can be interpreted as substituting of the entire set of multiple features with only one, single feature. This kind of approach can be recognized as the opposite to standard feature extraction or selection techniques, since it forms a feature set on the basis of one feature, while the existing feature extraction or selection methods determine the final feature set as a subset of the initial feature set. In our experiments in the field of the game of chess, we used, as a mentioned above quantity, the difference between the chess position score produced by the chess engine on two fixed depths of the decision tree search.

Feature Discovery with the Hellinger Distance

The main goal in the introduced solution is to initially separate two subsets in the entire analyzed data set using the Hellinger distance. The intention is to obtain two subsets, distinct in the space of some features. To accomplish this, one may take advantage of the statistical model involving probability distributions described in the previous section. The differences between objects in these subsets should be chosen as the features of the objects in the analyzed data set, since they will reveal discriminative information about the objects in the set, clearly indicating relevant features.

Let K denote the analyzed data set and let ρ_K denote the general discrete probability distribution associated with the set K. Building of the distribution ρ_K is described in the previous section. We intend to separate from the set K two subsets denoted as L and M, which we will call the Hellinger-distinct sets. The subsets L and M are characterized by the discrete distributions ρ_L and ρ_M respectively. Let denote the lower bound of the Hellinger distance and let β denote the upper bound of the Hellinger distance, which will be used to define the Hellinger-distinct sets L and M. Let γ be the cardinality of the set L and of the set M. Let MAX be the maximal number of iterations in the Procedure 1, before the interval pointed by the Hellinger distance bounds α and will be narrowed. This ensures that the Procedure 1 will terminate in a finite number of steps. The choice of parameters α, β, γ, and MAX is arbitrary. The values, we assumed in our experimental study, are given in section describing the experiments.

Definition 5 *The Hellinger-distinct sets L and M in the set K are defined as the sets satisfying following conditions:*

1. $L \subset K,\ M \subset K$.
2. $\mathrm{H}(\rho_K, \rho_L) \leq \alpha,\ \mathrm{H}(\rho_K, \rho_M) \geq \beta$.
3. $\alpha < \beta$.

The process of generating distributions ρ_L and ρ_M, satisfying condition 2 of Definition 5, and the process of creating sets L and M for settled α, β, γ, and MAX, satisfying condition 3 of Definition 5, is described in the following procedure.

Procedure 1 *The creation of the Hellinger-distinct sets L and M in the set K.*

Step 1. *Initially assign* $L \leftarrow \emptyset,\ M \leftarrow \emptyset,\ n \leftarrow 0$.

Step 2. *Final cardinality of the sets L and M settle to the value γ.*

Step 3. *Draw the object* $k \in K$.
$n \leftarrow n + 1$.
If $n \geq MAX$, *then* $\alpha \leftarrow \alpha + 0.01,\ \beta \leftarrow \beta - 0.01,\ n \leftarrow 0$.

Step 4. *If* $\mathrm{H}(\rho_{L\cup\{k\}}, \rho_K) \leq \alpha$ *and* $|L| < \gamma$, *then* $L \leftarrow L \cup \{k\}$,
else if $\mathrm{H}(\rho_{M\cup\{k\}}, \rho_K) \geq \beta$ *and* $|M| < \gamma$, *then* $M \leftarrow M \cup \{k\}$,
else go to Step 3.

Step 5. *If* $|L| < \gamma$ *or* $|M| < \gamma$, *then go to Step 3.*

Step 6. *The sets L and M are the sought sets.*

The distribution ρ_L is built by systematic appending randomly drawn samples, which do not increase the value of the Hellinger distance above the level α. Samples, which cause crossing this boundary, are rejected. Each sample is an object from the set K having assigned value of the chosen quantity associated with the set K. Consequently, one obtains the distribution ρ_L, for which the Hellinger distance from the distribution ρ_K is below the level α (ρ_L is near, in the meaning of the Hellinger distance, to the ρ_K). One also obtains a set L, from which the samples in the distribution ρ_L come. In the case of the distribution ρ_M, which should be far, in the meaning of the Hellinger distance, from the ρ_K distribution, one accepts only samples, which do not decrease the Hellinger distance below the level β during the distribution building process. And, as a result one gets the set M.

The Procedure 1 leads to creation of the subsets L and M of the analyzed data set K allowing to reveal the essential differences between the objects in the original set K. These differences should be used as features. At this point, it is clear how the concept proposed in this paper leads to effective feature discovery, since it is easier to determine features as the differences between the objects in two sets than choosing them without any additional knowledge about the data.

In our experiments, we employed the Hellinger distance defined as the square of H. Defined this way, it is not a metric (see section describing the Hellinger distance). However, our purpose was to show that despite it does not satisfy all conditions of the metric's definition, it is still a very useful tool in the feature discovery problem.

Experiments

In our experiments, we wanted to show, that separation of the Hellinger-distinct sets L and M in the data set K, can be helpful in determination of the most essential features of the objects in the set K. As a field of the experiments, we have chosen the game of chess, since it is a difficult and complex domain. As analyzed data set, we have used a set of chess positions of a specific material configuration, i.e., positions with white king, rook, knight, and pawn vs. black king, rook, knight, and pawn, called *KRNPkrnp* positions, for short. The example is discussed in more details in [20].

The random quantity associated with the set K, in the sense of the Definition 3, was the difference between score returned by the chess engine after 10 ply deep search and the 15 ply deep search. The quantity defined this way, may be considered as a random variable, since it is strongly dependent on some random factors, like contents of the hash table used during the game tree search process, hash table size, or unknown (to the user of the engine) search algorithm extensions. The score difference data gathered for some set

of samples into the distribution constitute the general discrete probability distribution in the sense of the Definition 4. In our case, the number of samples was 2000, and the resulting distribution is shown on Figure 1.

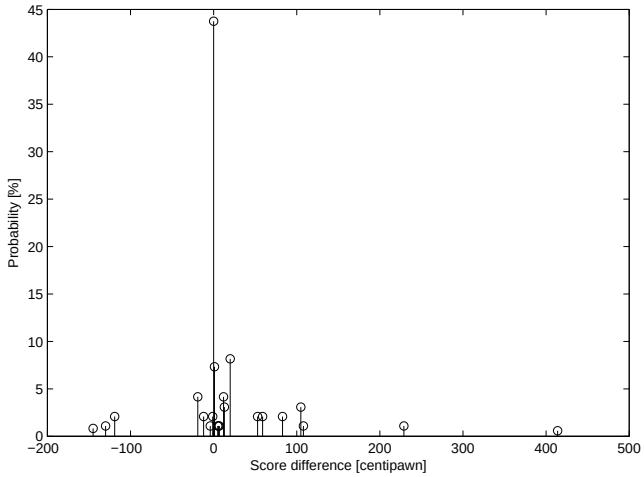

Fig. 1. A general discrete probability distribution (score difference distribution) for a set containing *KRNPkrnp* positions. A single pawn is worth 100 centipawns

For a problem formulated this way, we focus on the discovering of features of the *KRNPkrnp* chess positions, which is not a simple task, in general, even for an expert at grandmaster level. We give an example of the feature of the chess positions, discovered with our method. We also test the quality of separation of subsets L and M from the set K, by comparison with the results of some classical method of clustering applied to the set $L \cup M$. The experiment was carried out according to the following procedure:

Procedure 2 *Let the following data be given:*

- *the set K with the general discrete probability distribution ρ_K,*
- *any clustering method based on the features of the objects in the clustered data set.*

The procedure goes as described below.

Step 1. Separate the Hellinger-distinct sets L and M in the set K, according to the Procedure 1, for the fixed boundaries α and β $(\alpha < \beta)$, and fixed cardinalities of the sets L and M.

Step 2. Carry out the clustering of the set $L \cup M$ using any clustering method based on the features of the objects in the set $L \cup M$.

Step 3. Evaluate the obtained results by testing, whether the classical method generated at least one cluster containing only (or in majority) objects from the set L or from the set M.

Existence of at least one such cluster can be considered as a case confirming the hypothesis that the Hellinger distance can successfully suggest the division criterion for the sets L and M (by showing the differences between the objects of the sets L and M), which can be regarded as the discovered feature.

In the experiment, the following assumptions were stated:

- Let the set K be the set of all chess positions of type *KRNPkrnp*,
- Let ρ_K be the general discrete probability distribution associated with the set K,
- Let the quantity associated with the set K be the difference between the chess position score produced by the chess engine on depth 10 and 15 half-moves of the decision tree search,
- Let the clustering method used for evaluation of the obtained results be the expectation maximization (EM) method,
- Let the boundaries $\alpha = 0.2$, $\beta = 0.9$, and the cardinality $\gamma = 50$, for both sets L and M. Let $MAX = 1000$.

Following the Procedure 2, we get the following:

1. The Hellinger-distinct sets L and M in the set K, such that $\mathrm{H}(\rho_K, \rho_L) \leq 0.2,\ \mathrm{H}(\rho_K, \rho_M) \geq 0.9$,
   (a) the set L (having 50 objects) contains 20 positions with at least one pawn one line before the promotion rank (second rank for black and seventh rank for white),
   (b) the set M (having 50 objects) contains only 8 positions characterized by this feature,
2. EM-clustering results for the set $L \cup M$ with 10 clusters, among which, there is one particularly interesting cluster containing only the objects possessing feature described before (8 of 10 positions in this cluster come from the set L, see Figure 2).

Even after brief examination of the sets L and M it is easy to notice that a feature, which differs the objects of the set K well, is the ***possession of a pawn placed one line before the promotion rank, by at least one of sides***. Execution of the Procedure 2 results in discovering the feature described above. The clustering of the set $L \cup M$ with usage of the EM method confirmed this choice by forming a cluster, which contains only the objects characterized by this feature. It is worth to notice that the discovered feature is the binary-type feature, which would be difficult to detect by known feature extraction or feature selection algorithms. Figure 2 presents all positions from the best cluster, which suggests the above feature very clearly.

Open Problems

The problem which remains unsolved is the proper choice of the quantity associated with the set (Definition 3), which should be a random variable. This paper gives no detailed, principled way to find it, noting that the choice of it should not be arbitrary, since the further phases of the proposed approach strongly depend on this quantity. Therefore, it should reflect the most discriminative and relevant information about the objects in the analyzed data set. These are the requirements of a standard feature extraction or selection problem, however, the advantage of our method is that it needs only one feature, on the basis of which the whole feature set might be formed. This can be interpreted as substituting of the feature set with only one, single feature.

Summary

In this paper, we proposed an approach to the feature discovery based on the utilization of certain statistical means, specifically the Hellinger distance. The general concept was to separate two subsets (the Hellinger-distinct sets) from the analyzed data set in order to extract the differences between these subsets and use them as features of the objects in the analyzed set. In other words, the Hellinger-distinct sets allow to identify the highly discriminative division criteria for the original set, which are the desirable features.

During the experiments we carried out the comparison of the results obtained by the introduced procedure with the results of a classical feature-based EM clustering method. The EM method formed a cluster containing only the objects possessing the feature discovered by the presented algorithm. This confirms that the discovered feature provides a good

discrimination of the objects in the analyzed data set, and, therefore, is recommended for use on further stages of the analysis. We gave an example of the feature clearly suggested after the execution of the presented procedure. The mentioned feature is the binary-type feature, which is difficult to detect by the classical feature extraction or feature selection methods. The application of the Hellinger distance in the feature discovery problem can be considered as an essential improvement in the case of difficult and complex domains, like the game of chess, for example.

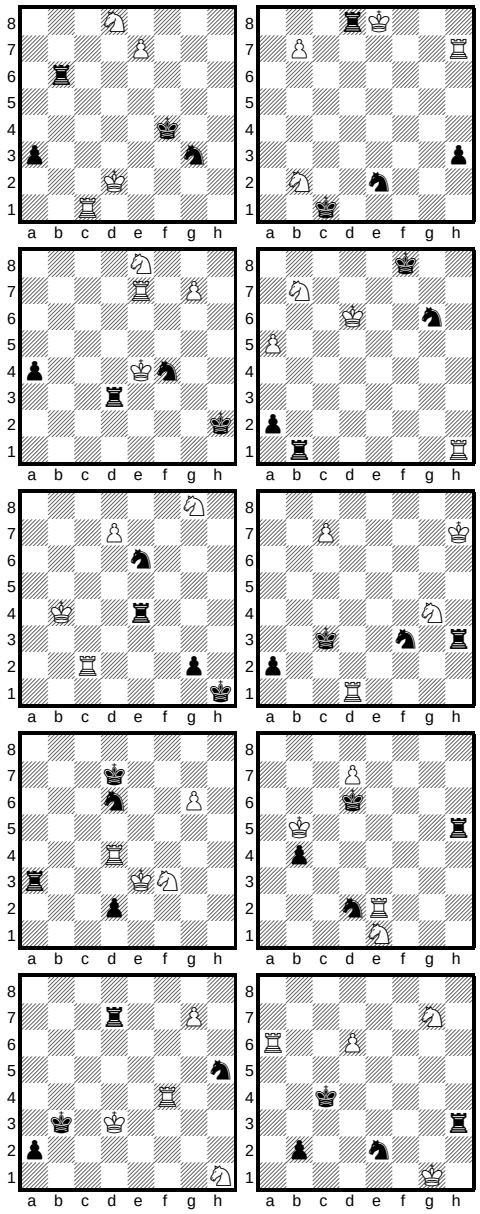

Fig. 2. One of the clusters obtained with the EM method. In each position, white is to move. Each position of this cluster has at least one pawn before the promotion rank

The proposed approach can be utilized as an independent feature discovery algorithm, however, it can be also considered as a data pre-processing tool supporting the existing approaches to either feature extraction or feature selection. In this case the Hellinger distance-based feature retrieval may serve as a procedure for generating the initial feature set on which the feature extraction or selection methods will be invoked. This will provide the initial feature set consisting of relevant features, which will be additionally processed by the feature extraction or selection algorithms allowing to obtain features being a result of combined, two-stage feature retrieval method.

BIBLIOGRAPHY

[1] B. C. Kuo and K. Y. Chang, "Feature Extraction for Small Sample Size Classification Problem," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 756–764, March 2007.

[2] S. F. Ding, Z. Z. Shi, Y. C. Wang, and S. S. Li, "A Novel Feature Extraction Algorithm," in *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, vol. 3. IEEE, August 2005, pp. 1762–1767.

[3] J. Mao and A. K. Jain, "Artificial Neural Networks for Feature Extraction and Multivariate Data Projection," *IEEE Transactions on Neural Networks*, vol. 6, no. 2, pp. 296–317, March 1995.

[4] E. Choi and C. Lee, "Optimizing Feature Extraction for Multiclass Problem," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 521–528, March 2001.

[5] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[6] B. Krishnapuram, A. J. Hartemink, L. Carin, and M. A. Figueiredo, "A Bayesian Approach to Joint Feature Selection and Classifier Design," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1105–1111, September 2004.

[7] H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithm for Classification and Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, April 2005.

[8] L. Wang, "Feature Selection with Kernel Class Separability," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1534–1546, September 2008.

[9] J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi, and Z. Chen, "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 3, pp. 320–333, March 2006.

[10] P. F. Hsieh, D. S. Wang, and C. W. Hsu, "A Linear Feature Extraction for Multiclass Classification Problems Based on Class Mean and Covariance Discriminant Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 223–235, February 2006.

[11] B. Bursteinas and J. Long, "Transforming Supervised Classifiers for Feature Extraction," in *Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence*. IEEE, November 2000, pp. 274–280.

[12] S. F. Ding, W. K. Jia, C.-Y. Su, and Z. Z. Shi, "Research of Pattern Feature Extraction and Selection," in *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics*, vol. 1. IEEE, July 2008, pp. 466–471.

[13] D. Olszewski, M. Kolodziej, and M. Twardy, "A Probabilistic Component for K-Means Algorithm and its Application to Sound Recognition," *Przeglad Elektrotechniczny*, vol. 86, no. 6, pp. 185–190, June 2010.

[14] M. Basseville, "Distance Measures for Signal Processing and Pattern Recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, December 1989.

[15] D. Pollard. (2000) Asymptopia. Book in progress. [Online]. Available: http://www.stat.yale.edu/ pollard/Books/Asymptopia/

[16] A. L. Gibbs and F. E. Su, "On Choosing and Bounding Probability Metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, February 2002.

[17] H. Abdi, "Distance," in *Encyclopedia of Measurement and Statistics*, N. J. Salkind, Ed. Sage Publications, Thousand Oaks, CA, 2007.

[18] A. A. Borovkov, *Mathematical Statistics*. Gordon and Breach Science Publishers, 1998.

[19] P. Diaconis and S. L. Zabell, "Updating Subjective Probability," *Journal of the American Statistical Association*, vol. 77, no. 380, pp. 822–830, December 1982.

[20] R. Lopatka and V. Rajlich, "On Feature Discovery in Board Games," in *Game-On Conference (GameOn'NA 5th Annual North American 2009)*, August 2009.

Authors: Rafał Łopatka, Ph. D., email: lopatkar@ee.pw.edu.pl, Dominik Olszewski, M. Sc., email: olszewsd@ee.pw.edu.pl, Faculty of Electrical Engineering, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warszawa, Poland.