ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The Choquet Kernel for Monotone Data

Ali Fallah Tehrani, Marc Strickert, and Eyke Hüllermeier

Computational Intelligence Group, Philipps Universität Marburg
D-35032 Marburg, Germany

Abstract. In this paper, we introduce a kernel for monotone data derived from the Choquet integral with its underlying fuzzy measure. While a naïve computation of this kernel has a complexity that is exponential in the number of data attributes, we propose a more efficient approach with quadratic time complexity. Kernel PCA and SVM classification are employed to illustrate characteristics and benefits of the new *Choquet kernel* in two experiments related to decision-making and pricing.

Keywords: Choquet integral, kernels, fuzzy measure, monotone data

1 Introduction

The Choquet integral is an aggregation operator that is commonly used in fields such as multi-criteria decision making and decision under uncertainties [6, 10]. Being based on a non-additive measure on the set of attributes, it allows for expressing positive and negative interactions between individual attributes (criteria). At the same time, the Choquet integral guarantees monotonicity, i.e., a monotone dependency of the aggregation on each individual attribute.

Generally, the problem of monotone classification and regression, in which higher attributes induce higher explanatory variables, has received increasing attention in neural networks [4, 12] decision tree learning [9] and ensemble models [3] in recent years. The aforementioned ability of the Choquet integral to model interactions between attributes is not limited to two-way interactions [8] but may comprise any number of attributes; complex data properties can thus be captured in a very flexible way [5]. The flexibility to account for relations between all subsets of attributes comes at the cost of an exponential complexity.

In this paper, we propose a new kernel based on the Choquet integral that can be computed in polynomial time. Equipped with this kernel, common spectral methods for clustering and classification can be applied to given datasets. We focus on support vector machines (SVM) and visualizations by means of kernel principal component analysis for illustration purposes.

2 The Discrete Choquet Integral

The Choquet integral is a non-additive aggregation function defined by an underlying fuzzy measure. Let $C = \{c_1, \ldots, c_n\}$ be a finite set and μ a measure $2^C \to [0, 1]$. For each $A \subseteq C$, we interpret $\mu(A)$ as the weight of the set of elements A. Additivity is a standard assumption on a measure $\mu(\cdot)$, that is, $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \subseteq C$ such that $A \cap B = \emptyset$. Unfortunately,

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

such measures cannot model interaction between elements, because the extension of a set of elements $A \subseteq C$ by a set of elements $B \subseteq C \setminus A$ always increases the weight $\mu(A)$ by the weight $\mu(B)$, regardless of A and B.

As opposed to this, fuzzy measures can be non-additive and be used for modeling higher-order interactions between attributes. Monotonicity is reflected by

$$\mu(\emptyset) = 0, \mu(C) = 1 \quad \text{and} \quad \mu(A) \leq \mu(B) \;\; \text{for all} \;\; A \subseteq B \subseteq C \; . \tag{1}$$

The *Möbius transform* is a useful representation of non-additive measures:

$$\mu(B) = \sum_{A \subseteq B} \boldsymbol{m}(A) \;\; \text{with} \;\; \boldsymbol{m}(A) = \sum_{E \subseteq A} (-1)^{|A|-|E|} \cdot \mu(E) \;\; \text{for all} \;\; B \subseteq C \; . \tag{2}$$

$\boldsymbol{m}(A)$ can be interpreted as the weight that is *exclusively* assigned to A, instead of being indirectly connected with A through interaction with other subsets.

If $c_i \in C$ are considered as binary attributes encoding for presence or absence, $\mu(A)$ can be seen as an integral of the indicator function f_A of A given by $f_A(c) = 1$ if $c \in A$ and $= 0$ otherwise. Alternatively, $f \colon C \to \mathbb{R}^+$ can be any non-negative function that assigns a value to each criterion c_i. For example, $f(c_i)$ might be the degree to which a criterion c_i is satisfied. The aggregation of function values $f(c_i)$ into an overall evaluation, weighted according to the measure μ, can be expressed by an integral $\mathcal{C}_\mu(f)$ of f w.r.t. to the measure μ. Choquet defined this integral for an underlying fuzzy measure as [1]

$$\mathcal{C}_\mu(f) = \sum_{i=1}^{n} \big(f(c_{(i)}) - f(c_{(i-1)})\big) \cdot \mu(A_{(i)}) \; . \tag{3}$$

The notation $(\cdot)$ refers to a permutation of $\{1, \ldots, n\}$ such that $0 \leq f(c_{(1)}) \leq f(c_{(2)}) \leq \cdots \leq f(c_{(n)})$ (with $f(c_{(0)}) := 0$), and $A_{(i)} = \{c_{(i)}, \ldots, c_{(n)}\}$. Using the Möbius transform of μ, the Choquet integral can be rewritten as follows:

$$\begin{aligned} \mathcal{C}_\mu(f) \; &= \; \sum_{i=1}^{n} f(c_{(i)}) \cdot \big(\mu(A_{(i)}) - \mu(A_{(i+1)})\big) \; = \; \sum_{i=1}^{n} f(c_{(i)}) \sum_{R \subseteq T_{(i)}} \boldsymbol{m}(R) \\ &= \sum_{T \subseteq C} \boldsymbol{m}(T) \cdot \min_{i \in T} f(c_i) \;\; \text{with} \;\; T_{(i)} = \big\{S \cup \{(i)\} \,|\, S \subset \{(i+1), \ldots, (n)\}\big\} \; . \end{aligned} \tag{4}$$

The Möbius transform is a useful means to derive a Choquet kernel representation, because it allows for expressing the basis functions of the Choquet integral in a monotonic way. We can rewrite the expression (4) as inner product

$$\mathcal{C}_\mu(f) = \sum_{T \subseteq C} \boldsymbol{m}(T) \cdot \min_{i \in T} f(c_i) = \Big\langle \boldsymbol{m}_\mathfrak{I}, \varphi\big(f(\boldsymbol{x})\big) \Big\rangle \; . \tag{5}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The feature mapping function $\varphi : \mathbb{R}^n \to \mathbb{R}^{2^n-1}$ and $\boldsymbol{m}_{\mathfrak{I}}$ are defined as

$$\varphi(\boldsymbol{x}) = \varphi(x_1, ..., x_n) = \Big(x_1, ..., x_n, \min\{x_1, x_2\}, ..., \min\{x_{n-1}, x_n\}, \min\{x_1, x_2, x_3\}, ..., \min\{x_1, ..., x_n\}\Big) , \tag{6}$$

$$\boldsymbol{m}_{\mathfrak{I}} = \Big(\boldsymbol{m}\big(\{c_1\}\big),...,\boldsymbol{m}\big(\{c_n\}\big),\boldsymbol{m}\big(\{c_1, c_2\}\big),...,\boldsymbol{m}\big(\{c_{n-1}, c_n\}\big),...,\boldsymbol{m}\big(\{c_1,..., c_n\}\big)\Big).$$

Thus, the explicit feature mapping of n-dimensional input vectors $\boldsymbol{x}$ implements a minimization over all $2^n - 1$ subsets of attributes.

3 Choquet Kernel

The feature mapping is used to define an inner product between $\varphi(\boldsymbol{x})$ and $\varphi(\boldsymbol{x^*})$:

$$\Big\langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{x^*}) \Big\rangle = x_1 x_1^* + \ldots + x_n x_n^* + \min\{x_1, x_2\} \min\{x_1^*, x_2^*\} + \ldots + \min\{x_1, x_2, \ldots, x_n\} \min\{x_1^*, x_2^*, \ldots, x_n^*\}$$

We write $\big\langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{x^*}) \big\rangle = \sum_{T \subseteq \{1,\ldots,n\}} \min_{i \in T}\{x_i\} \cdot \min_{i \in T}\{x_i^*\}$ as $K_{\mathcal{C}}(\boldsymbol{x}, \boldsymbol{x}^*)$ and note that this summation gives rise to a valid kernel.

Explicit computation of the Choquet kernel $K_{\mathcal{C}}(\boldsymbol{x}, \boldsymbol{x}^*)$ would involve $2^n - 1$ summands which is not very practical for higher-dimensional input vectors. By reformulating the evaluation, the complexity can be reduced to $\mathcal{O}(n^2)$. We exploit the fact that the kernel expression is invariant under permutation, i.e.,

$$\big\langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}^*) \big\rangle = \big\langle \varphi(\sigma(\boldsymbol{x})), \varphi(\sigma(\boldsymbol{x}^*)) \big\rangle = \big\langle \boldsymbol{x}, \boldsymbol{x}^* \big\rangle + \sum_{i=1}^{n-1} x_{\sigma_i} \cdot \left\{ \sum_{s=1}^{n-i} \wp(i, s) \right\} ,$$

$$\text{with} \quad \wp(i, s) = \underbrace{\sum_{j=i+1}^{n} \sum_{k=j+1}^{n} \ldots \sum_{p=o+1}^{n}}_{s\text{-summations}} \min \left\{ x_{\sigma_i}^*, x_{\sigma_j}^*, x_{\sigma_k}^*, \ldots, x_{\sigma_p}^* \right\} . \tag{7}$$

The permutation σ describes the ordering $x_{\sigma_1} \leq ... \leq x_{\sigma_n}$ of attributes of $\boldsymbol{x}$. Let ξ be a permutation of an ordered subset $x_{\xi_1} \leq \ldots \leq x_{\xi_p}$ of p elements, then

$$\sum_{T \subseteq \{x_{\xi_1}, \ldots, x_{\xi_p}\}} \min\{T\} = \sum_{i=0}^{p-1} 2^{p-1-i} \cdot x_{\xi_{i+1}} . \tag{8}$$

By using the fact that for any y, $\min\big\{y, T\big\} = \min\big\{y, \min\{T\}\big\}$, we obtain

$$\sum_{\emptyset \neq T \subseteq \{x_{\xi_1}^*, \ldots, x_{\xi_p}^*\}} \min\{y, T\} = \sum_{j=0}^{p-1} 2^{p-1-j} \cdot \min\{y, x_{\Psi_j^*}^*\} . \tag{9}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Note that the left hand side in (9) is invariant under permutation, and in the right-hand side $x^*_{\Psi^*_j}$ is the j-th ordered value among $\{x^*_{\xi_1}, \ldots, x^*_{\xi_p}\}$. Putting everything together, an efficient formulation of the Choquet kernel is obtained:

$$K_{\mathcal{C}}(\boldsymbol{x}, \boldsymbol{x}^*) = \left\langle \boldsymbol{x}, \boldsymbol{x}^* \right\rangle + \sum_{i=1}^{n-1} x_{\sigma_i} \cdot \left\{ \sum_{j=0}^{n-1-i} 2^{n-1-i-j} \cdot \min \left\{ x^*_{\sigma_i}, x^*_{\Psi^*_{i,j+1}} \right\} \right\} . \quad (10)$$

This involves a nested summation over weighted minima of permuted elements and contains *contributions of all subsets of attributes*. Since these permutations are related to the prior ordering of attributes in $\boldsymbol{x}$ and $\boldsymbol{x}^*$, creating a complexity of $\mathcal{O}(n \cdot \log n)$, this sums up to an overall quadratic time complexity.

4 Applications

The **DenBosch database** [2] contains descriptions of 120 houses in the city of Den Bosch (NL) by 8 attributes: district, area, number of bedrooms, type of house, volume, storeys, type of garden, and number of garages. The output is a binary variable indicating whether the price of the house is low (61 instances) or high (59 instances), depending on whether or not it exceeds a threshold.

The Choquet kernel is compared with the popular RBF kernel and the polynomial kernel. While the feature space of RBF is infinitely dimensional, the polynomial order is set to 8 to account for all $2^8 - 1$ attribute interactions. This makes polynomial and Choquet kernels structurally comparable. Optimum SVM parameters were identified by five-fold nested cross validation. The average 0/1-loss over 20 separate runs using the Choquet kernel on testing data was 11.96%±6.89, 12.83%±4.98 for the RBF kernel and 22.60%±8.17 for the polynomial kernel. Compared to RBF, the better average results of the Choquet kernel come at the price of more unstable learning, indicated by larger standard deviation.

Figure 1 shows the different arrangements of points in the feature spaces of SVM models for the different kernels. The corresponding Gram matrices were turned into 2-D scatter plots by using kernel PCA [11] with first and second eigenvectors of the double-centered kernel-matrix shown as x- and y-axis, respectively. Although a substantial loss of information can be assumed for the 2-D embeddings compared to the original data relationships, a structural difference between the plots can be observed. The RBF kernel (left) spreads points quite well, but locally mixes up both classes. In contrast, the poly-8 kernel (center) exhibits an uneven point distribution with a wide scatter of high price houses. Class information for the Choquet kernel (right) coincides well with the x-axes, if the horizontal scale is properly taken into account.

Since, in the current model, no explicit monotonicity constraints are used, we cannot ensure the monotonicity of the measure extracted from the Choquet kernel. Therefore, we propose an indicator of the degree of monotonicity:

$$D_\mu = \frac{\left|\{(A,B) | A \subseteq B,\, |A| = |B| - 1,\, \mu(A) \leq \mu(B)\}\right|}{n \cdot 2^n} \in [0, 1] . \quad (11)$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

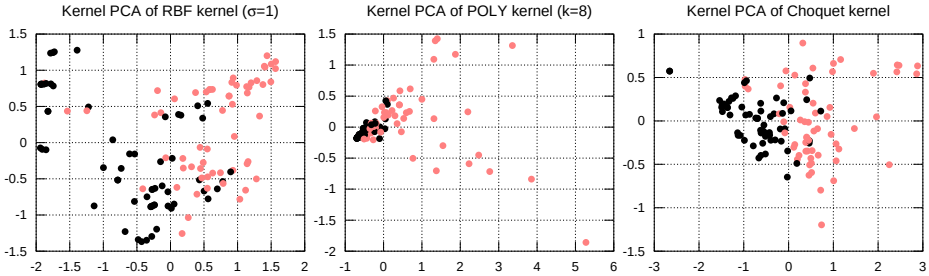

Figure 1: Kernel PCA of DenBosch house data with RBF kernel (left), Choquet kernel (center), and polynomial kernel (right). Black markers refer to low prices, red (gray) points to high prices.

If all monotonicity constraints are fulfilled, i.e. $D_\mu = 1$, μ is a fuzzy measure. Here, the 5-fold cross validation led to an average of $D_\mu^{80\%} = 0.886$. Thus, the calculated Choquet kernel does not perfectly represent the Choquet integral. Yet, it performs fairly well.

The **WEKA employee selection database** [7] contains profiles of applicants for industrial jobs. Four input attributes contain ordinal assessments of psychometric test results from candidate interviews. The output is an overall score on an ordinal scale between 1 and 9, corresponding to the degree of suitability of each candidate to this type of job. For binary classification, 248 suitable candidates (score 6–9) are to be distinguished from 240 unsuitable (score 1–5) subjects. Again, five-fold nested cross validation was carried out. The average 0/1-loss test error over 20 separate runs using the Choquet kernel was 5.10%±1.91, 10.41%±7.21 for a poly-4 kernel, and 7.11%±2.78 for the RBF kernel. The average degree of monotonicity in this 5-fold cross validation training is $D_\mu^{80\%} = 0.9936$. Thus, almost all monotonicity constraints are fulfilled, which explains the excellent behavior of the Choquet kernel.

5 Discussion and Conclusions

The parameter-free *Choquet kernel* and its efficient realization has been introduced for analysing monotone data. Two applications illustrate specific feature mapping properties of the new kernel: ordering relationships among attributes can be naturally accounted for in learning scenarios such as pricing and survey data. The structure of the proposed Choquet kernel is quite different from RBF kernels. Like polynomial kernels with order of data dimensionality, higher-order interactions between attributes are modeled, but without similarly strong overfitting at high orders. As shown for the employee data, Choquet kernels are designed to model monotonicity. The current work properly represents the Choquet integral for full monotonicity with D_μ close to 1. To cope with data like the DenBosch set with $D_\mu = 0.886$ or with more general data, upcoming research seeks for correction strategies of the fuzzy measure to induce monotone kernels.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

So far, n-additivity is modeled in the Choquet formalism; this means that interactions from all possible subsets are taken into account. To decrease the complexity of this huge internal dependence structure, the cardinality of modeled attribute subsets could be restricted to $k < n$, i.e., to k-additive Choquet integrals. In this case, the proposed all-subset simplification does no longer work. While 2-additivity can be expressed as $K_{\mathcal{C}}^{k=2}(\boldsymbol{x}, \boldsymbol{x}^*) = \langle \boldsymbol{x}, \boldsymbol{x}^* \rangle + \sum_{i<j} \min(x_i, x_j) \cdot \min(x_i^*, x_j^*)$, larger $k < n$ involve much more complex terms and longer run times. Low-order additivity might better reflect weak dependencies between attributes in real-life data sets. Thus, automatic order estimation as well as related performance studies point into important directions of future work. Programs are online available at `https://mloss.org/software/view/537`

Acknowledgments: This work was supported by the German Research Foundation (DFG) and by the Marburg Research Center for Synthetic Microbiology (SYNMIKRO), funded within the LOEWE program by the state Hesse.

References

[1] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1954.

[2] H. Daniels and B. Kamp. Applications of MLP networks to bond rating and house pricing. *Neural Computation and Applications*, 8:226–234, 1999.

[3] K. Dembczyński, W. Kotlowski, and R. Slowinski. Learning rule ensembles for ordinal classification with monotonicity constraints. *Fund. Inform.*, 94(2):163–178, 2009.

[4] W. Duivesteijn and A. Feelders. Nearest neighbour classification with monotonicity constraints. In *Machine Learning and Knowledge Discovery in Databases*, volume 5211 of *Lecture Notes in Computer Science*, pages 301–316. Springer, 2008.

[5] A. Fallah Tehrani, W. Cheng, K. Dembczyński, and E. Hüllermeier. Learning monotone nonlinear models using the Choquet integral. In *Machine Learning and Knowledge Discovery in Databases*, volume 6913 of *Lecture Notes in Computer Science*, pages 414–429. Springer Berlin Heidelberg, 2011.

[6] M. Grabisch and C. Labreuche. Fuzzy measures and integrals in MCDA. In J. Figueira, S. Greco, and M. Ehrgott, editors, *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 563–608. Springer Verlag, Boston, Dordrecht, London, 2005.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[8] E. Hüllermeier and A. Fallah Tehrani. Efficient learning of classifiers based on the 2-additive Choquet integral. In C. Moewes and A. Nürnberger, editors, *Computational Intelligence in Intelligent Data Analysis*, volume 445 of *Studies in Computational Intelligence*, pages 17–29. Springer Berlin Heidelberg, 2013.

[9] R. Potharst and A. Feelders. Classification trees for problems with monotonicity constraints. *ACM SIGKDD Explorations Newsletter*, 4(1):1–10, 2002.

[10] Y. Rébillé. Decision making over necessity measures through the Choquet integral criterion. *Fuzzy Sets and Systems*, 157(23):3025–3039, 2006.

[11] B. Schölkopf, A. Smola, and K. Müller. Kernel principal component analysis. In *Advances in kernel methods: Support vector learning*, pages 327–352. MIT Press, 1999.

[12] J. Sill. Monotonic networks. In *Advances in Neural Information Processing Systems*, pages 661–667. The MIT Press, Denver, USA, 1998.