Depth based support vector classifiers to detect data nests of rare events

Rainer Dyckerhoff and Hartmut Jakob Stenz

January 30, 2018

Extended Abstract

The aim of this project is to combine the idea of data depth with support vector machines (svm) for classification. To this end, we introduce data depth functions and svm and discuss why a combination of the two is assumed to work better in some cases than using svm alone. Given a sample or "cloud" X of data points $x_1, \ldots, x_n$ in $\mathbb{R}^d$, a data depth is a function $\mathcal{D}(z|X) : \mathbb{R}^d \to [0,1]$ that describes how "deep" z lies in X: Values near to 0 mean z is far away from the center of X and values close to 1 mean the opposite, so z is located in or next to the center of X. Put more formally, every function $\mathcal{D}(z|X)$ which fulfills the following axioms is called data depth function:

- **Affine invariant:** $\mathcal{D}(Az + b|AX + b) = \mathcal{D}(z|X)$ for all $b \in \mathbb{R}^d$ and bijective linear transformation $A : \mathbb{R}^d \to \mathbb{R}^d$.
- **Null at infinity:** $\lim_{||z|| \to \infty} \mathcal{D}(z|X) = 0$.
- **Monotone on rays:** For z^* with $\mathcal{D}(z^*|X) = \max_{z \in \mathbb{R}^d} \mathcal{D}(z|X)$ and any $\alpha > 0$ and r out of the unit sphere S^{d-1}, $\mathcal{D}(z^* + \alpha r|X)$ decreases in a weak sense.
- **Upper semi continuous:** The upper level sets $\mathcal{D}_\alpha(X) := \{z \in \mathbb{R}^d : \mathcal{D}(z|X) \geq \alpha\}$ are closed for all α.

The relevant literature consists of many research contributions relating to data depths. The different approaches range from the development of individual depths such as the "Mahalanobis depth" (Mahalanobis, 1936, see [6]) or the "halfspace depth" (Donoho *et al.*, 1992, see [1]) over the formalization and systematization of certain depth characteristics (e.g. Dyckerhoff, 2004, see [2]) to using data depths for classification purposes. Concerning the latter approach, the work of Li *et al.* (2012, see [5]), which developed the *DD*-classification as a new classification method based on data depths, is worth mentioning: For a unkown function $f : \mathbb{R}^d \to \{-1, 1\}$ one can define classes $X = \{z_i \in \mathbb{R}^d | f(z_i) = -1\}$ and $Y = \{z_j \in \mathbb{R}^d | f(z_j) = 1\}$ and ask for a individual whether it should assinged to class X or Y, a typical binary classification problem.

Li *et al.* analyze the set $\{(x,y) \in \mathbb{R}^2 | x = \mathcal{D}(z_k|X), y = \mathcal{D}(z_k|Y), k = 1, \ldots, n\} \subset [0,1]^2$, the so called *DD*-Plot, for given training points $z_1, \ldots, z_n$. The aim is to construct a decision rule based on the *DD*-Plot instead of using the initial data points for reference. After determining the depths $x^* = \mathcal{D}(z^*|X)$ and $y^* = \mathcal{D}(z^*|Y)$, a new data point z^* can thus be classified by applying this decision rule to (x^*, y^*). In recent years, scientists took a number of different approaches as to how a decision rule based on the *DD*-Plot can be constructed: While Li *et al.* (2012) applied polynomial dividing lines, Mozharovskyi (2014, see [7]) among others uses a non-parametric approach in his *DD*-α-procedure. This project aims to take up the approach of Kim *et al.* (2017, see [4]). It is based on the application of svm in the *DD*-Plot meaning that the different points of the *DD*-Plot are separated with the help of separating hyperplanes employing kernels. In doing so, one expects that the use of svm will allow for the construction of a decision rule that will produce precise results even in the event of overlapping cases. In addition, the relevant literature offers highly diverse and detailed articles discussing the mentioned classification tool (e.g. Friedman *et al.*, 2001, see [3]).

The focus of this project lies on the detection of rare events, which are structured in data nests: Class X contains much more data points than class Y and Y has less dispersal than X. Therefore the data cloud of Y is located in the data cloud of X. Such data structures play a significant role in churn prediction analyses (see Reuß and Zwiesler, 2004, [8]), for example. Therefore, the application described in such publications serves as a potential but not sole motivation to analyze the central issue of how to predict the likelihood of a customer's churn given that the training set contains merely a very limited number of churn customers in comparison to "normal" customers. It is very well possible that such churn customers deviate on average only very little from others but that they are nevertheless clustered in certain segments. As a consequence, there is a difference in dispersal rather than in location of the set X (normal customers) and Y (churn customers). As there additionally is a disequilibrium of the two data sets, this form of classification problem is akin to finding the proverbial needle in a haystack.

From an analytical point of view, this project strives to demonstrate that under the assumption of a multivariate normal-distribution the separability in the *DD*-Plot will increase with increasing data dimension if and only if the characteristics of data nests are fulfilled and the data clouds strongly overlap. It may seem intuitive to assume that access to more information will allow for a more precise classification. However, the observation that the structure of the data itself seems to be the reason that makes this classification possible deserves further investigation: Due to applying a data depth transformation and transferring the structure into the *DD*-Plot, the number of possible dimensions is drastically reduced from d to merely two. What is more, this transformation is irreversible. Yet, the question whether important information is actually lost or rather compressed when carrying out this process remains. Ultimately, using the *DD*-Plot results in a radical simplification of the classification problem. Or put differently, the result is a trade-off between a loss of information and a reduction of complexity of the problem. The latter aspect in particular is of great significance in a big data context.

Beyond the analytical investigations, comprehensive simulation studies will be carried out. The goal is to analyze randomly distributed data in different nest structures and to test different depths for the construction of the *DD*-Plot as well as various kernels for the svm. This might ultimately allow for a generalization of the analytical findings. On top of that, a method comparison is part of the project as well: Both with regard to simple svm without data depth transformation as well as in relation to other classification methods, the application of the described approach on different data examples is intended to illustrate whether said approach works better than other approaches. It is important to note that this approach does not require sampling, which is commonly used when analyzing rare events. Last but not least, the project will expand on the difficulty of using particular depths in the context of high dimensional data and discuss possible solutions to this problem.

References

[1] Donoho, D. L. and Gasko, M. [1992], 'Breakdown properties of location estimates based on halfspace depth and projected outlyingness', *The Annals of Statistics* pp. 1803–1827.

[2] Dyckerhoff, R. [2004], 'Data depths satisfying the projection property', *Allgemeines Statistisches Archiv* **88**, 163–190.

[3] Friedman, J., Hastie, T. and Tibshirani, R. [2001], *The elements of statistical learning*, Vol. 1, Springer New York.

[4] Kim, S., Mun, B. M. and Bae, S. J. [2017], 'Data depth based support vector machines for predicting corporate bankruptcy', *Applied Intelligence* pp. 1–14.

[5] Li, J., Cuesta-Albertos, J. A. and Liu, R. Y. [2012], 'Dd-classifier: Nonparametric classification procedure based on dd-plot', *Journal of the American Statistical Association* **107**, 737–753.

[6] Mahalanobis, P. C. [1936], 'On the generalised distance in statistics', *Proceedings of the National Institute of Sciences of India, 1936* **12**, 49–55.

[7] Mozharovskyi, P. [2014], *Contributions to depth-based classification and computation of the Tukey depth*, Dr. Kovac Verlag.

[8] Reuß, A. and Zwiesler, H.-J. [2004], Stornoanalyse in einem unfallversicherungsbestand, *in* 'Proceedings der 8. Konferenz der SAS®-Anwender in Forschung und Entwicklung (KSFE), Aachen', pp. 265–276.