**Use Case 4: Classification with sparse genetic data - OV**

In the use case 3, netDx is used to build a predictor from sparse data such as DNA CNVs data. Instead, this example demonstrates how to use netDx with somatic DNA mutations, sparse data type which can be handled with the approach developed by Hofree at al. for his method NBS [1]. Combining prior knowledge of gene interactions with network propagation is possible to decrease the sparsity. Here we build a case/control classifier for the Ovarian serous cystadenocarcinoma (OV) downloaded from TCGA [2] data portal (<https://tcga-data.nci.nih.gov/tcga>), starting from point mutations (single nucleotide polymorphism as well as di/tri/oligo-nucleotide polymorphism) and indels of patients who survived to the disease against those who did not. Silent mutations were filtered out and mutations profiles were defined as binary vectors with ones whenever a patient is mutated in a given gene and zeros otherwise.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pan-cancer tumor dataset** | **Dimension**  **(genes x patients)** | **Mutated genes**  **per patient** | **Total Percentage of mutations in the dataset** | **Deceased vs**  **Living** |
| Ovarian carcinoma | 10192 x 363 | 0,005% | 0,5% | 172 vs 191 |

**Propagation approach**

This use case introduces a strategy to reduce the sparsity of the genetic data based on the approach developed by Hofree at al. in NBS. The application of the method requires the usage of a gene-gene interaction network. As default, netDx uses the cancer specific network developed for the python implementation of NBS [3]. This network is defined by a set of nodes which are genes and a set of edges such that an edge connects two nodes representing their physical direction interaction. It is a compact cancer reference network (CRN) that contains only high-confidence interactions specific to cancer. It has been proved that the CRN effectively clusters tumor samples from several different cancer types, as measured by the clusters’ ability to predict patient survival, in comparison to one of the networks used in the original NBS study. First, netDx removes the genes not represented as nodes in the CRN from the patient data. Second, it processes each patient profile individually. The nodes of the CRN representing patient’s a priori mutated genes assume a value equal to 1 while all the others a value equal to 0. This numerical information is propagated through the network following the guilt by association principle [Fig. A,B] Specifically, the algorithm applied for the propagation is the random walk with restart which has been the most selected in bioinformatics applications [1,3,4] and implemented by Jonathan Ronen at al. in R [7]. Nodes closer to the a priori mutated genes will tend to assume a value higher than those nodes that are more distant. The result of using this strategy on a patient’s binary somatic mutation profile is a non-sparse network-smoothed profile in which the state of each gene is a continuous score that reflects its network proximity to the a priori mutations and how much it is involved to them. Finally, netDx applies a discretization. It sets to 1 those genes having a high propagation score, while zero in the opposite case. Precisely, it sets to 1 only a number of genes which is equal to 1.3 multiplied by the number of original mutations. This value has been determined empirically testing the approach on 8 different cancer datasets [Fig. C]. It proved to be a valid choice in term of performances and guarantees to select only the best 1% scored genes [Supplementary: Table 1] (lower is the score and lower is the association to a priori mutations). At the end, netDx gets a less sparse somatic mutation data matrix to build the predictor.

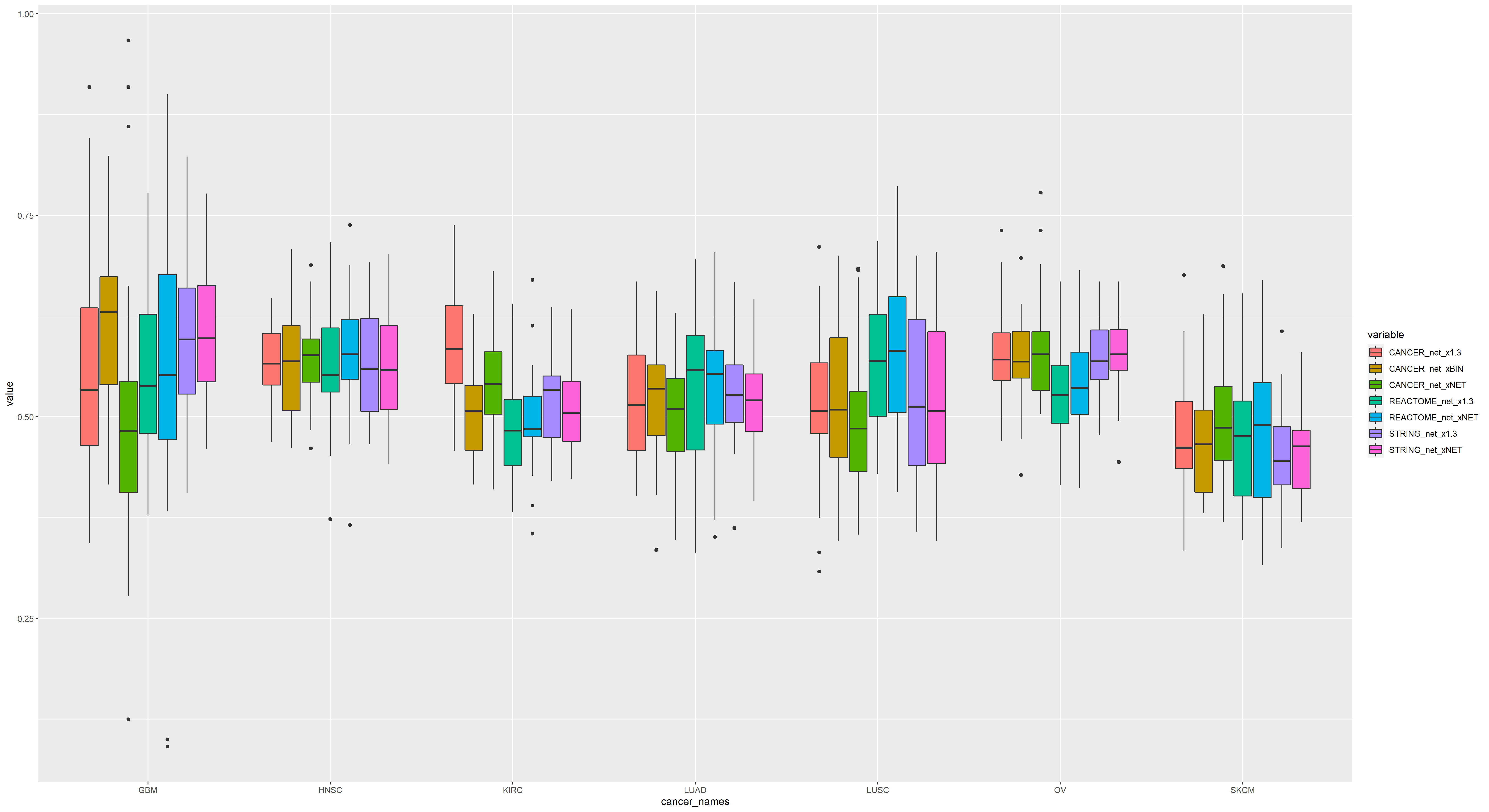
Fig. A) it shows the propagation approach used to prepare the somatic mutation data. Binary (1 black, 0 white) patient profiles compose the input matrix provided to netDx. The software applies the propagation individually to each profile and exploiting a gene-gene interaction network. It ends binarizing the resulting matrix of scores assigned by the propagation to each gene. Only the top scored genes get a 1 and are considered having an imputed mutation. Fig. B) It shows the effect of the propagation on one mutation of a blue patient. Close genes get a propagation score, distant genes get a score close or equal to zero.

Fig. C) netDx prediction performances using the propagation approach on 8 different cancer datasets. It has been applied with 3 different networks (STRING, REACTOME and cancer specific). Plus, we compared the final performances with the ones got with and without specific steps. The suffix “1.3” indicates that netDx applied the propagation approach with the 1.3 multiplier. The suffix “NET” indicates that netDx did not apply the propagation, but it built the predictor using the input matrix filtered of genes not existing in the network. The suffix “BIN” indicates that netDx built the predictor with the input binary matrix and the binary similarity function.

**Design and Adapting the Algorithm for Sparse Event Data**

In this design, we group mutations by pathways. The reason behind is that cancer is a disease not of individual mutations, nor of single genes, but of combinations of genes acting in molecular networks corresponding to hallmark processes such as cell proliferation and apoptosis[5,6]. In addition, we include in the grouping also imputed mutations derived by the propagation approach. One a priori non-mutated gene is considered mutated if it is selected in the discretization step due to its score. In this case, the logic has been proved by Hofree at al. The researchers have been able to cluster smoothed somatic mutation profiles of three major cancers into robust tumor subtypes having a strong association to the survival time. Proving that the propagation scores are informative to find markers which act on specific biological processes of patients.

**Binary Similarity and Label enrichment**

In this design, the similarity used is still the binary function. If two patients share a mutation in a pathway, their similarity for that pathway is one; otherwise it is zero. On the contrary of the approach for CNV data, the label-enrichment step is not applied. This because the sparsity is mitigated with the propagation which, as also indicated for the grouping, enhances the number of patient profiles sharing a pathway.

**Cumulative feature scoring** and **Evaluating model performance** are sections which remained unchanged with respect their description proposed in the use case 3.

[1] Hofree M, Shen JP, Carter H, Gross A, Ideker T, Network-based stratification of tumor mutations, 2013.

[2] The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20. pmid:24071849

[3] Justin K Huang, Tongqiu Jia, Daniel E Carlin, Trey Ideker, pyNBS: a Python implementation for network-based stratification of tumor mutations, *Bioinformatics*, Volume 34, Issue 16, 15 August 2018

[4] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **6**, e1000641 (2010).

[5] Kreeger, P.K. & Lauffenburger, D.A. Cancer systems biology: a network modeling perspective. Carcinogenesis 31, 2–8 (2010).

[6] Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. Cell 144, 646–674 (2011).

[7] Ronen J and Akalin A. *netSmooth:* Network-smoothing based imputation for single cell RNA-seq [version 2; peer review: 2 approved]. *F1000Research* 2018, **7**:8 (<https://doi.org/10.12688/f1000research.13511.2>)