

Local machine learning-based feature importances for gene regulatory network inference

Auteur : Kerff, Alexandre

Promoteur(s) : Geurts, Pierre; Huynh-Thu, Vân Anh

Faculté : Faculté des Sciences appliquées

Diplôme : Master : ingénieur civil en informatique, à finalité spécialisée en "management"

Année académique : 2023-2024

URI/URL : <http://hdl.handle.net/2268.2/21141>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

Local machine learning-based feature importances for gene regulatory network inference

Alexandre Kerff - Master in Computer engineering

2023 - 2024

Supervisor : Pr. Pierre Geurts

Understanding how a cell (or organism) reacts to a change in the environment or disturbance requires an understanding of the intricate processes controlling gene expression and, therefore, protein synthesis. A common representation of these mechanisms is the gene regulatory network, that aims at defining the regulation links between genes as a set of interactions. Inferring those gene regulatory networks from expression data has been a widely studied field at the level of bulk expression data. However, recent breakthroughs in sequencing technologies enables measurements at the resolution of a single cell. Such data allows the development of research towards the analysis of gene regulatory networks for a single specific cell or for a distinct cell type, rather than global interactions. This thesis has the objective to perform these analyses.

Exploiting the foundations of a technique elaborated for bulk data, **Genie3**, the problem of cell-specific and cell-type specific network inference can be addressed by the means of local feature importance methods instead of global algorithms.

To this extent, this study first examines numerous local feature importance methods and provides new implementations for a few of them. It evaluates them with respect to the global methods on simple regression problems. Analysing these techniques highlights particular methods of interest with promising results (**Shap**, Saabas, local mean decrease of impurity, and local mean decrease of accuracy).

Subsequently, the local methods are employed to address the cell-specific network inference issue in order to examine their applicability in this domain. To evaluate the anticipated local networks' capacity to identify distinct interactions, they are contrasted with global networks on a synthetic dataset. It is demonstrated that analysing the local feature importance algorithms (**Shap**, Saabas, local mean decrease of impurity, and local mean decrease of accuracy) yields more accurate findings at single-cell resolution than analysing the global networks.

Next, the efficient local methods are investigated in the context of cell-type specific network inference. This problem is addressed in the thesis by averaging the local scores of cells sharing the same types on a synthetic dataset with mixed types. Comparing the methodology to the application of global method to each separated type, it is shown that all the local algorithms perform poorly.

Subsequently, a real dataset containing gene expression data from several cell types obtained from peripheral blood is subjected to local techniques. As there are no ground truth networks accessible, the cell-type feature importance inference is performed to recover rankings of features scores. Common and unique interactions between types are highlighted by comparing the most important significance values for each type. The use of local mean decrease of impurity is shown to identify common and different rankings than global methods run on the separated datasets.

Lastly, certain genetic markers found in each patient that contributes to the actual dataset are examined. A correlation is calculated between networks created from patients who have the genetic marker and patients who do not. A few low correlation values found by local mean decrease of impurity enable the identification of certain indicators that affect gene regulation.

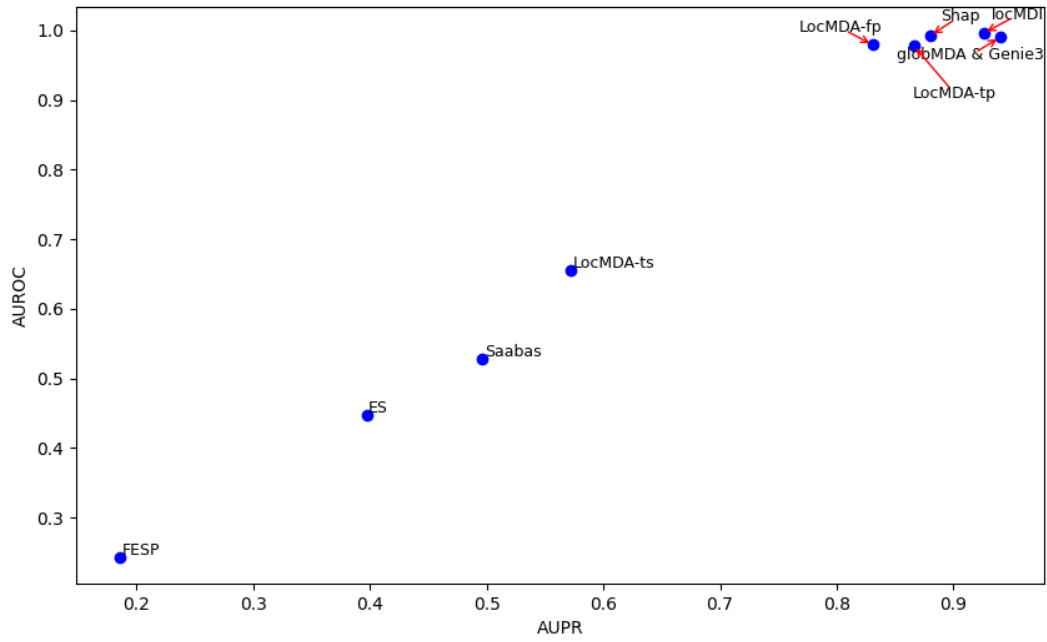


Figure 1: *mean of meanAUROC/meanAUPR plot for local methods and global methods on Friedman 1 dataset*

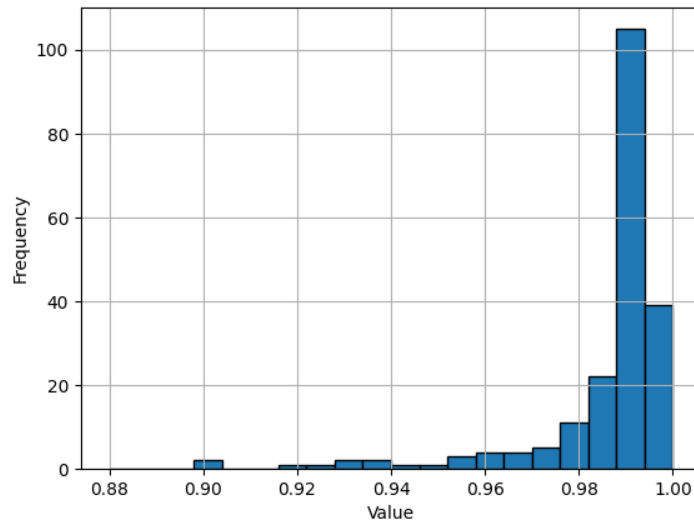


Figure 2: *Histogram of correlation values between feature importance values for the absence and presence of each genetic markers*

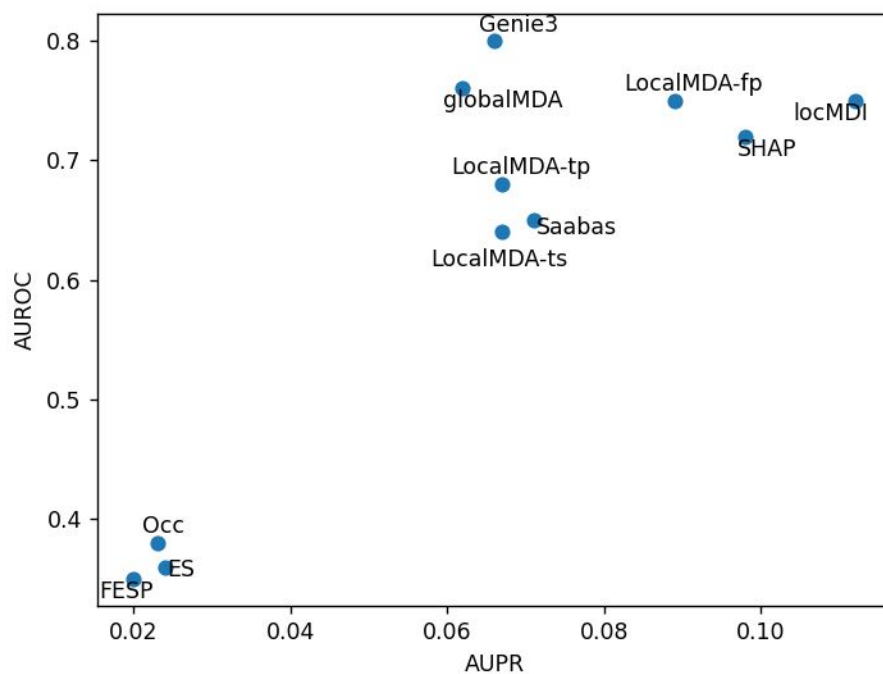


Figure 3: *mean of meanAUROC/meanAUPR plot for best normalized local methods and global methods for Dyngen datasets*