Feature selection in machine learning

Classification is one of the essential tasks in machine learning whose purpose is to classify each instance in the dataset into different classes based on its features. It is often difficult to determine which features are useful without prior knowledge. As a result, a large number of features are usually introduced into the data set that may be irrelevant or redundant. Feature selection is process of selecting small subset of relevant features from original large set of features. This small subset of features may have less redundant or relevant features making the machine learning process simple with reduced learning process time and increased performance. Other benefits of feature selection are improved prediction performance, scalability, understandability, and generalization capability of the classifier. It also reduces computational complexity and storage, provides faster and more cost-effective model, and knowledge discovery. Moreover, it offers new insights for determining the most relevant or informative features. The main challenge that occurs in feature selection is large search space where for n datasets, solutions is 2^n. Feature selection consists of complex stages that usually are costly. And even the optimal model parameters of full feature set might need to be redefined for a few times in order to obtain the optimal model parameters for selected feature subsets. Feature selection also involves two main objectives, which are to maximize the classification accuracy and minimize the number of features, which are both conflicting objectives. Hence, feature selection is considered as multi-objective problem with some trade-off solutions that lie in between these two objectives. Some examples of feature selection techniques are Information Gain, chi-square, lasso and Fisher Score. Feature selection can be used to find key genes (i.e., biomarkers) from a large number of candidate genes in biological and biomedical problems, to discover core indicators or features to describe the dynamic business environment, to select key terms like words or phrases in text mining and to choose or construct important visual contents like pixel, color, texture, and shape in image analysis. In comparison to other dimensionality reduction techniques such as those based on projection for example, principal component analysis (PCA) or compression, feature selection techniques do not modify the original representation of the variables, but simply select a subset of them. Hence, they maintain the original semantics of the variables offering interpretability.

Feature selection used on gene expression data which has small sample size is called gene selection. Gene selection can be used to find key genes from biological and biochemical problems. This type of feature selection is important for disease detection and discovery such as tumor detection and cancer discovery which results in giving better diagnosis and treatment. Gene expression data can be expressed as fully labelled, unlabeled, or partially labelled. This leads to development of supervised, unsupervised and semi-supervised gene selection to discover biological patterns and classes. There are many feature selection methods such as

supervised, unsupervised and semi-supervised feature selection. In Supervised feature selection, it uses the labelled data for feature evaluation. But the data is large and continues to collect data in increasing rate. Moreover, the labelled data is costly to obtain and may be undependable and mislabeled which may cause over-fitting the learning process in supervised type feature selection by either removing relevant features or using irrelevant features. In the case of supervised method, previous knowledge is taken into account. Unsupervised feature selection is more difficult to work with than other two approaches because it is unaided by labelled data. But advantages of this type of feature selection are unbiased and perform well with no previous knowledge. Unsupervised feature selection is useful in discovery of disease and classification of disease types. The disadvantage of unsupervised approach is it ignores connection between different features and it depends on some mathematical principles with no guarantee that those principles are valid for all data. Semi-supervised feature selection is combination of supervised and unsupervised feature selection. Semi-supervised feature selection is also being used for gene classification by jointly employing both labelled and unlabeled data.

Gene expression data can be evaluated using microarray data methods is essential with different samples. These methods can be grouped into unsupervised, supervised and semi-supervised methods. The microarray data has large number of genes which are redundant. Thus, it needs to identify some important genes for better understanding of the fundamental data, also minimize the time taken for improved post-processing tasks such as classification, subset selection of genes (features) and so on. Using Feature selection, a subset of relevant features can be selected from the original large set of features. For finding key genes from large number of from a large number of applicant genes in biological and biomedical problems using features like genes, biomarkers and so on. Biomarker is a feature which gives indication of medical condition observed from the patient externally and this can be measured as well as reproducible and different than medical symptoms which show only the signs regarding disease or health that are understood only by the patients themselves. Feature selection has several advantages for microarray data.

- First, dimension reduction to reduce the computational cost.
- Second, reduction of noises to improve the classification accuracy.
- Finally, more interpretable features or characteristics that can be helpful to identify and monitor the target diseases.

Biologically, only a few genetic alterations correspond to the malignant transformation of a cell. Determination of these regions from microarray data can allow for high-resolution global gene expression analysis to genes in these regions and better biological problem detection and classification for better diagnosis, prognosis and correct treatment for corresponding biological problems.