Feature Embedding improves Clustering of Breast Cancer Microarray Data

Christina Morgenstern
christinamorgenstern@lewisu.edu
DATA-51000-002, 20
Data Mining and Analytics
Lewis University

I. Introduction

Through the application of new technologies in the life sciences, a tremendous amount of data has been and still is being generated. This so-called *omics*-revolution has produced e.g. genomic, transcriptomic or proteomic datasets by the application of high-throughput technologies yielding information of the molecular make-up of an organism. Gene expression data contains the information of which genes are switched on or off as well as the magnitude of their expression in different cellular context. Comparative studies between different cell types or between healthy and cancerous cells are performed to infer knowledge in the process of cell differentiation and tumorigenesis, respectively [1].

Cancer denotes a diverse group of genetic diseases and accounts for the second leading cause of death in the world according to the World Health Organization (WHO) [2]. Experiments using microarrays are able to capture differences in thousands of genes simultaneously and are used in cancer research to infer changes in gene expression levels between healthy and cancerous cells. While this data can provide valuable insights into the process of tumor development, cancer prognosis and treatment the analysis thereof has proved challenging in the past. Recently, machine learning algorithms (ML) have evolved and become applicable to microarray data for classification and clustering tasks [3].

The goal of this study is to improve clustering on a breast cancer microarray dataset to serve as future benchmark for identifying and characterizing clusters of genes that play an important role in tumorigenesis.

In the following sections, a description of the dataset used is given (II) as well as the methodology laid out (III). The results are presented and discussed in chapter IV and a conclusion in part V summarizes the efforts of this paper.

II. Data Description

The dataset used in this work contains breast cancer gene expression data from microarray analysis and was downloaded from Kaggle [4]. The data was curated and provided by the Curated Microarray Database (CuMiDa), a database that contains pre-processed microarray datasets for machine learning projects [5]. The CuMiDa repository contains 78 handpicked datasets that have undergone normalization, background correction, sample viability, sample quality analysis and personalized editing to yield reliable datasets [5]. The dataset with the identifier GSE45827 contains breast cancer gene expression data from 151 samples and 54,676 genes. This results in the "large p, small n problem" representing a large number of dimensions (i.e. the gene expression values) and only a small sample size. The problem is also known as *The Curse of Dimensionality* which can lead to problems in clustering and classification tasks [3].

The microarray expression profiling was performed using samples from 41 patients with primary invasive breast cancer as well as from 11 healthy patients and 14 cell lines. The breast cancer samples are further divided into three classes: basal, HER2, Luminal A and Luminal B referring to the molecular subtypes of breast cancer. Briefly, ribonucleic acid (RNA) was extracted from these samples and the whole transcriptome analyzed using an Affymetrix U133 Plus 2.0 Chip [6]. This type of microarray covers the entire human genome and allows for the analysis of over 47,000 transcripts.

An overview of the data table used for the clustering analysis can be seen in Table 1 with the feature columns samples, cancer type and two examples of genes analyzed for expression using microarray experiments. The full description of the 54,675 features (genes) in the dataset can be accessed via the Gene Expression Omnibus (GEO) website [7]. Due to space limitations it cannot be listed here. In total, there are 151 unique identifiers (samples) and 54,677 features (gene expression values of 54,675 genes, a sample identifier and the type of breast cancer) in the breast cancer microarray dataset used in this study.

TABLE 1. DESCRIPTION OF DATASET

Attribute	Type	Example Value	Description
SAMPLES	Numeric (integer)	84	Record identifier
TYPE	Nominal (string)	“basal”	Type of invasive breast cancer
1007_s_at	Numeric (float)	9.850040	U48705 /FEATURE=mRNA /DEFINITION=HSU48705 Human receptor tyrosine kinase DDR gene, complete cds
1053_at	Numeric (float)	8.097927	M87338 /FEATURE= /DEFINITION=HUMA1SBU Human replication factor C, 40-kDa subunit (A1) mRNA, complete cds

No missing values were present in the data as stated by the authors and verified in the analysis. For the exploratory data analysis und subsequent clustering approaches, the data was loaded into the Jupyter notebook [8] environment and analyzed using Python 3 programming software [9]. The following libraries were used: NumPy [10], pandas [11], Matplotlib [12], seaborn [13] and scikit-learn [14].

The dataset contains samples from patients with different breast cancer types (basal, HER, Luminal_A and Luminal_B) as well as samples from healthy patients (normal) and data from cell line experiments. The distribution of these labels can be seen in Fig. 1. The majority of labels with 41 samples account for the basal-like breast cancer type, a particularly aggressive molecular subtype defined by genes expressed by epithelial cells in the mammary gland [15]. 30 samples are of breast cancer type HER, referring to the growth-promoting protein HER2 [16]. Luminal A and Luminal B breast cancer subtypes refer to the origin of the tumor cells in the luminal (inner) cells lining the mammary ducts [17]. These four types are the most common molecular subtypes in breast cancer diagnosis and are characterized by a distinct pattern of gene expression [18].

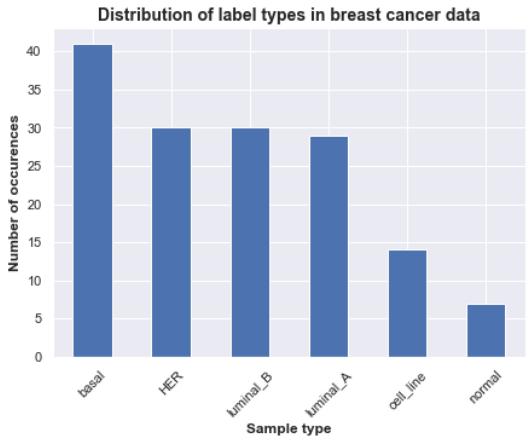

Fig. 1. Distribution of label types in breast cancer dataset.

Using descriptive statistics, the feature columns were explored for their distribution in means, standard deviation, minimum and maximum values and different percentile measures. All the feature values are float values and range between approx. 0 and 20. A sample of the columns is shown in Table 2. According to the authors of the CuMiDa database, the features have already been normalized [5], which can be verified by looking at the data.

TABLE 2. DESCRIPTIVE STATISTICS ON FEATURE COLUMNS (EXAMPLE)

	Samples	1007_s_at	1053_at
count	151.000000	151.000000	151.000000
mean	160.668874	10.338901	7.631910
std	45.431226	0.613445	0.706464
min	84.000000	7.505488	5.855968
25%	121.500000	10.103030	7.166075
50%	159.000000	10.416819	7.531673
75%	200.500000	10.735117	8.053832
max	238.000000	11.675109	9.627008

III. METHODOLOGY

Analysis of microarray data using clustering approaches can reveal hidden patterns in gene expression data such as which genes can drive tumorigenesis. The approaches to clustering microarray data are either based on genes but also on samples or subspaces [19]. The focus in this study is to cluster samples based on the gene expression values.

Before applying the clustering techniques, the data was scaled such that each feature has unit variance. To achieve this, the StandardScaler module of the Scikit-learn machine learning library was used [14].

In this study different feature embeddings were used in order to improve the clustering benchmark published by CuMiDa. Using the algorithm k-means, the authors could achieve an accuracy score of 0.7 [5]. The flowchart in Fig. 2 show the steps taken in this work to improve clustering using the k-means algorithm and different feature embeddings. k-means is one of the most popular clustering algorithms that aims at minimizing within-cluster variances while maximizing between cluster variances [20]. Initially, the original dataset was taken and subjected to k-means clustering while subsequent approaches used the embeddings t-distributed stochastic neighbor embedding (t-SNE) [21] and uniform manifold approximation and projection (UMAP) [22] before applying the k-means algorithm. Each method was evaluated using silhouette plots and silhouette scores as well as an accuracy score and the rand-index. The results and the comparison of the different clustering approaches are presented in IV.

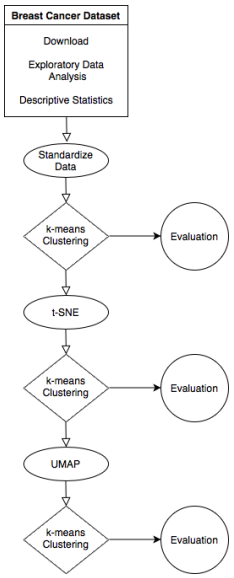

Fig. 2. Flowchart describing the approach on improving the clustering of the breast cancer microarray dataset.

IV. Results and Discussion

In this section, the results of the different clustering strategies are compared in terms of their accuracy of assigning a sample to the cluster with its known type label. Since in k-means clustering the number of k clusters must be known in advance, clustering was initially performed for a range of clusters from 1 to 20 to make sure the optimal number of k is used in the algorithm.

A. Determining the optimal number of clusters (k)

Although the number of clusters in the breast cancer dataset is already known with 6 types of cancer labels present, the cluster errors are calculated and plotted to for visual inspection of optimal k. In Fig. 3, the elbow plot is shown with the cluster errors plotted against the number of clusters. The k, at which the error stops to drop significantly (where the graph makes an elbow), determines the optimal number of k. The plot confirms that k=6 is indeed a good value for k-means clustering. Thus, in subsequent clusterings k was always set to 6.

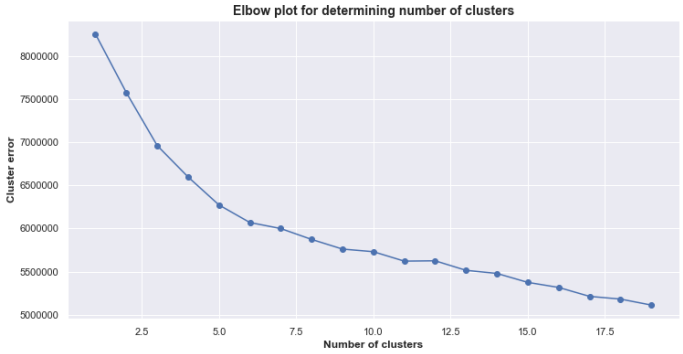

Fig. 3. Elbow plot of cluster errors against number of clusters.

B. Visualization of the high-dimensional breast cancer microarray dataset in 2D space

Feature embedding refers to the technique of translating high-dimensional data to low-dimensional space in order to improve ML algorithms while preserving the internal structure. In this study, the embeddings t-SNE and UMAP are used to improve k-means clustering on the breast cancer microarray dataset as compared to k-means clustering without embedding.

Embeddings also allow for visualization of high-dimensional data in two-dimensional space as seen in Fig. 4 which shows the data samples in 2D space after transformation with t-SNE and UMAP and labelled according to cancer type. The visual comparison of the two techniques shows that the t-SNE algorithm (Fig. 4a) does slightly better than the UMAP algorithm (Fig. 4b) in separating the individual types. Both algorithms are clustering the cell line (Fig. 4, green dots) and normal (Fig. 4, red dots) samples well apart from the other four cancerous sample types suggesting that these have markedly different gene expression profiles.

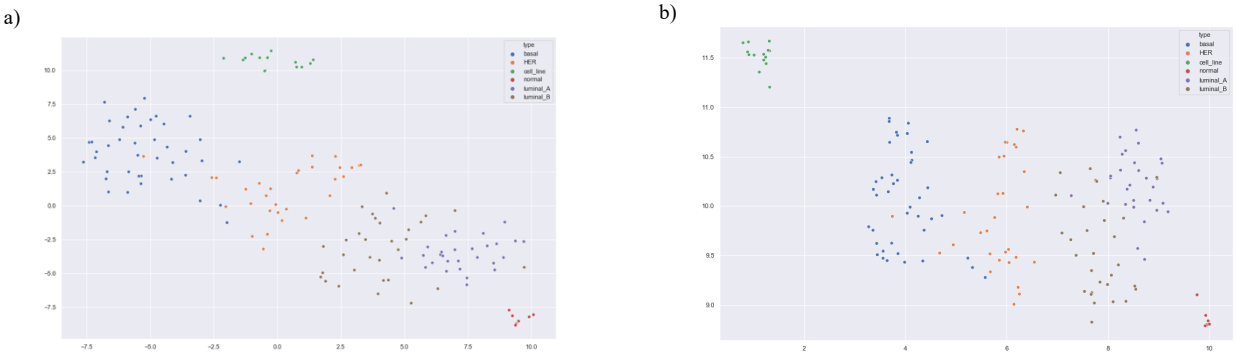

Fig. 4. 2D plot of breast cancer dataset using t-SNE (a) and UMAP (b) embeddings.

C. *Comparison of k-means clustering with and without embedding*

Accuracy, silhouette score and rand score are three different metrics used in this study to compare the clustering results. The accuracy score is the fraction of samples that have been assigned to the correct cluster, in our case the cluster label refers to the type of the sample [23]. It ranges between 0 and 1 and denotes bad and perfect cluster assignment. The silhouette score states the similarity of an object to other objects in the same cluster and how well the object separates from other clusters [24]. It can take values between -1 and 1 indicating poor cluster match and well match, respectively. The silhouette scores can also be visualized using silhouette plots. Lastly, the rand score considers all pairs of samples and measures the similarity between two clusterings by counting all pairs assigned in the same or different clusters [25]. A rand score of 0 denotes random clustering results while a rand score of 1 stands for perfect clustering.

The comparison of the k-means clustering results with and without embedding on the breast cancer microarray dataset using the above-mentioned metrics can be seen in Table 3. Without an embedding, the result of the k-means clustering is only 62% as measured by the accuracy score. Applying the k-means algorithm to the previously t-SNE embedded breast cancer microarray data increases the accuracy to 87%. The best accuracy with 94% is achieved when subjecting the data to UMAP-embedding before k-means clustering (Table 3).

TABLE 3. COMPARISON OF K-MEANS CLUSTERING ON BREAST CANCER DATASET

Embedding	Accuracy Score	Silhouette Score	Rand Score
No embedding	0.62	0.06	0.39
t-SNE	0.87	0.46	0.72
UMAP	0.94	0.56	0.85

The silhouette plots of the three clustering experiments further visualize the resulting silhouette scores for the individual clusters (see Fig. 5).

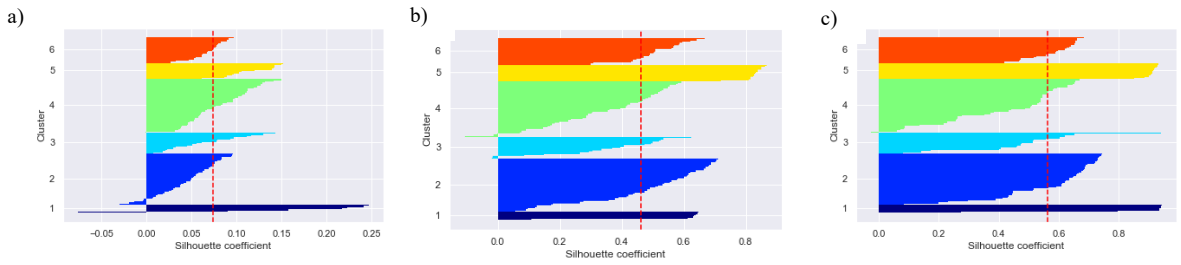

Fig. 5. Silhouette plots of k-means clustering without (a) and with t-SNE (a) and UMAP (b) embeddings.

V. CONCLUSIONS

Microarray data is inherent multi-dimensional with thousands of gene expression values and a comparatively small number of sample sizes [3]. This limits traditional clustering approaches and often leads to poor results. Since clustering of gene expression data is a promising approach for defining clusters of genes responsible for driving tumorigenesis it is vital to improve the clustering on this kind of data. In this study the curated microarray dataset containing 151 samples from breast cancer patients as well as healthy patients and cell lines was used with the goal to advance the application of k-means clustering [5]. Therefore, clustering was performed using the original dataset as well as applied to an embedded dataset using t-SNE and UMAP embeddings. The comparison of these three approaches demonstrated that embedding is a crucial step in clustering microarray data and improves the clustering results. Also, the previously published benchmark for k-means clustering on this dataset has been beaten using and embedding approach. The UMAP embedding increased the accuracy of the k-means clustering to 94% as compared to the published 70% [5].

Future studies will involve the application and verification of the mentioned strategy to other types of microarray data.

REFERENCES

[1] E. Mathé, J. L. Hays, D. G. Stover, and J. L. Chen, "The Omics Revolution Continues: The Maturation of High-Throughput Biological Data Sources," *Yearb. Med. Inform.*, vol. 27, no. 1, pp. 211–222, Aug. 2018, doi: 10.1055/s-0038-1667085.

[2] "Cancer." https://www.who.int/health-topics/cancer#tab=tab_1 (accessed Jun. 07, 2020).

[3] B. I. Grisci, B. C. Feltes, and M. Dorn, "Neuroevolution as a tool for microarray gene expression pattern identification in cancer research," *J. Biomed. Inform.*, vol. 89, pp. 122–133, Jan. 2019, doi: 10.1016/j.jbi.2018.11.013.

[4] "Breast cancer gene expression - CuMiDa." https://kaggle.com/brunogrisci/breast-cancer-gene-expression-cumida (accessed Jun. 07, 2020).

[5] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, "CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research," *J. Comput. Biol.*, vol. 26, no. 4, pp. 376–386, Apr. 2019, doi: 10.1089/cmb.2018.0238.

[6] T. Gruosso *et al.*, "Chronic oxidative stress promotes H2 AX protein degradation and enhances chemosensitivity in breast cancer patients," *EMBO Mol. Med.*, vol. 8, no. 5, pp. 527–549, May 2016, doi: 10.15252/emmm.201505891.

[7] "GEO Accession viewer." https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570 (accessed Jun. 07, 2020).

[8] F. Perez and B. E. Granger, "IPython: A System for Interactive Scientific Computing," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 21–29, 2007, doi: 10.1109/MCSE.2007.53.

[9] Pilgrim, M., & Willison, S. (, "Dive Into Python 3," vol. Springer, .

[10] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, Mar. 2011, doi: 10.1109/MCSE.2011.37.

[11] W. McKinney, "Data Structures for Statistical Computing in Python," presented at the Python in Science Conference, Austin, Texas, 2010, pp. 56–61, doi: 10.25080/Majora-92bf1922-00a.

[12] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May 2007, doi: 10.1109/MCSE.2007.55.

[13] Michael Waskom *et al.*, *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo, 2017.

[14] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Mach. Learn. PYTHON*, p. 6.

[15] D. J. Toft and V. L. Cryns, "Minireview: Basal-Like Breast Cancer: From Molecular Profiles to Targeted Therapies," *Mol. Endocrinol.*, vol. 25, no. 2, pp. 199–211, Feb. 2011, doi: 10.1210/me.2010-0164.

[16] T. Ishikawa *et al.*, "The role of HER-2 in Breast Cancer," *J. Surg. Sci.*, vol. 2, no. 1, pp. 4–9, Dec. 2014.

[17] Z. Inic *et al.*, "Difference between Luminal A and Luminal B Subtypes According to Ki-67, Tumor Size, and Progesterone Receptor Negativity Providing Prognostic Information," *Clin. Med. Insights Oncol.*, vol. 8, pp. 107–111, Sep. 2014, doi: 10.4137/CMO.S18006.

[18] "Gene Expression Profiling Identifies Molecular Subtypes of Inflammatory Breast Cancer | Cancer Research." https://cancerres.aacrjournals.org/content/65/6/2170 (accessed Jun. 06, 2020).

[19] S. Srivastava and N. Joshi, "Clustering Techniques Analysis for Microarray Data," p. 6, 2014.

[20] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.

[21] L. van der Maaten and G. Hinton, *Visualizing data using t-SNE*. 2008.

[22] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *ArXiv180203426 Cs Stat*, Dec. 2018, Accessed: Jun. 07, 2020. [Online]. Available: http://arxiv.org/abs/1802.03426.

[23] "sklearn.metrics.accuracy_score — scikit-learn 0.23.1 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html (accessed Jun. 07, 2020).

[24] "sklearn.metrics.silhouette_score — scikit-learn 0.23.1 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html (accessed Jun. 07, 2020).

[25] "sklearn.metrics.adjusted_rand_score — scikit-learn 0.23.1 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html (accessed Jun. 07, 2020).