Exploring conditional missing patterns for automated bacteria identification using MALDI-TOF MS data

J.C.F. da Rocha[1], A.Campos Jr.[1], R.M. Etto[2], C.W. Galvão[3]
G.L. Fedacz[4], R.R. da Silva[4], A.S.S. Oliveira[4]

[1]DEINFO/LIC – Universidade Estadual de Ponta Grossa (UEPG)
Av. Gen. Carlos Cavalcanti, 4748 – 84.030-900 – Ponta Grossa – PR - Brazil

[2]DEQUIM/LABMON – Universidade Estadual de Ponta Grossa

[3]DEBIOGEN/LABMON – Universidade Estadual de Ponta Grossa

[4]PPGCA – Universidade Estadual de Ponta Grossa

jrocha@uepg, arion@uepg.br, rmetto@uepg.br, carolinawgalvao@hotmail.com

Abstract. *Training classifiers for automated bacterial identification using MALDI-TOF fingerprints requires addressing class-conditional missingness patterns (CMPs). A CMP is a non-missing-at-random pattern that provides evidence for classification. One possible strategy to handle CMPs is feature stratification. This work evaluated the effectiveness of stratification in training naive Bayes classifiers for the proposed task through two experiments. The first experiment compared the predictive performance of categorical naive Bayes classifiers trained on stratified/discretized features with the performance of a Gaussian naive Bayes fitted on imputed data. The second experiment assessed the impact of class imbalance on the differences in the performance of Gaussian and categorical naive Bayes classifiers. The ANOVA results suggest that feature stratification can induce more accurate classifiers. Correlation analysis shows that class imbalance has a low influence on the difference in the performances of classifiers.*

Resumo. *A aprendizagem de classificadores para identificação automática de bactérias a partir* fingerprints *de espectrometria MALDI-TOF requer o tratamento de conjuntos de dados incompletos cuja ausência dos dados é condicional à hipótese de classificação (CMP). CMP é um padrão de perda não-aleatória (MNAR) que fornece evidencias para classificação. Uma estratégia para tratar o CMP é aplicar a estratificação de características. Considerando isto, este trabalho avaliou a eficácia da estratificação no treinamento de classificadores naive Bayes com a realização de dois experimentos. O primeiro, comparou o desempenho preditivo de classificadores categóricos, treinados sobre dados*

estratificados, com o desempenho de classificadores Gaussianos treinados em dados previamente imputados. O segundo experimento estimou o impacto do desbalanceamento de classe na diferença dos desempenhos dos classificadores Gaussianos e categóricos. Os resultados da ANOVA sugere que a estratificação de características induz a aprendizagem de classificadores mais acurados. A análise de correlção mostrou que o desbalanceamento de classes teve pouca influência sobre a diferença no desempenho dos classificadores.

1. Introduction

The use of software tools to identify microorganisms using Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry with Time-of-Flight data (MALDI-TOF MS) fingerprints is a promising technology for performing tasks in modern agriculture and food industry (Ashfaq et al., 2022) (Haider et al., 2023). For example, the detection of plant growth-promoting bacteria in a soil sample or the identification of pathogens related to plant diseases in sustainable agriculture (de Souza et al., 2015). These methods also can be used to reduce the risk associated with food contamination in the agroindustry (Zhang et al., 2022).

A core step in developing such tools is the supervised learning of a classifier. An algorithm that implements a probabilistic, logical, or functional model, enabling discrimination of the taxonomy of a microorganism based on its spectral fingerprint. To achieve this, the data analysis team must select a classification function from a wide range of ones. Next, they must train this function using an incomplete dataset. This is a challenging task, and despite this, the system users have high expectations of the classifier's performance.

The MALDI-TOF MS fingerprint datasets utilized for training classifiers in bacterial identification exhibit a specific missing data pattern known as class conditional missingness patterns (CMP)(Lin and Haug, 2008) (Ke et al., 2022). CMP is a *missing not-at-random* (MNAR) pattern that arises when the targets (the objects to be classified) do not express a given feature. Consequently, the presence or absence of such features provides evidence for the classification hypotheses. In the context of the proposed application, the CMP indicates that the microorganism does not express some ribosomal proteins in its spectra.

Value imputation, data removal, and estimation-maximization-based procedures are mainstream strategies for handling incomplete data in machine learning (Dong and Peng, 2013). However, they do not address the CMP patterns. An alternative approach is to apply a technique called feature stratification (Lin and Haug, 2008). This is a two-step procedure. First, a new constant element is appended to the domain of each feature that shows the CMP pattern. Next, each missing value is replaced with a constant; *nob* (not observed).

Stratification assumes discrete/categorical features. Naive Bayes addresses this prerequisite because this is a common preprocessing step in its application (Yang and Webb, 2009). Additionally, as noted in Tahan and Asadi (2018), discretization can mitigate the impact of class imbalance on supervised learning. That is a common condition in bioinformatics (Zhang et al., 2020).

This work presents the results of two experiments conducted to evaluate the effec-

tiveness of stratification when learning naïve Bayes classifiers for bacterial identification using MALDI-TOF fingerprints. In the first experiment, the predictive performance of a categorical naive Bayes trained on data with stratified features was compared to that of a Gaussian naive Bayes classifier trained on imputed data. The second experiment analyzed the correlation between class imbalance and the difference in the performance of categorical and Gaussian naive Bayes. The experiments utilized eight datasets. Two rates of missing data were considered: 20% and 30%. A constrained search-based and a mixed integer linear programming-based procedure were used to discretize the data. Balanced accuracy was used to assess the classifier performance.

An ANOVA test with *post hoc* analysis showed that feature stratification could induce more accurate classifiers. More objectively, the difference in the performance of the categorical and Gaussian classifiers favors the categorical classifiers and is often statistically significant. Correlation analysis showed that class imbalance did not degrade the performance gain obtained by feature stratification.

2. Background review

Machine learning has proven to be an efficient approach in developing tools for automated microorganism identification using peptide mass fingerprints (PMF) obtained by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight mass spectrometry (MALDI-TOF MS). The basic idea is to use supervised machine learning procedures to train classifiers on class-labeled datasets. Various classifiers have been used for carrying out that task (Weis et al., 2020). The Ribopeaks software (Tomachewski. et al., 2018) implemented a bacterium species identification based on the naive Bayes classifier (Mitchell, 1997).

The naive Bayes classifier utilizes the Bayes theorem to compute the posterior probability of each classification hypothesis. It can be viewed as a Bayesian network (Mitchell, 1997) whose dependence map is an n-ary tree of height equal to one. The root node represents the target variable, C, with domain Ω. Ω lists every classification hypothesis $c_1 \ldots, c_t$. A leaf node denotes a feature, X_i, $i = 1..n$, whose domain is Ω_i. If Ω_i is categorical/continuous, the respective node stores a table with the conditional distribution/density $p(X_i|C)$. The root node stores the marginal distribution $p(C) = (P(c_1) \ldots, P(c_t))$.

In this work, Ω is defined as $\{true, false\}$. When $C = true$, it indicates the microorganism in the sample belongs to the taxonomy (species, genus, family, etc) targeted by the classifier. Otherwise, $C = false$. Each leaf node corresponds to a unique entry in the PMF pattern. A PMF pattern is a tuple $\mathbf{x}$ with n elements. The $_i$-esim element stores the abscissa of the spectral peak related to a ribosomal protein expressed in the sample.

An evidence $\mathbf{x} = (x_{1,k_1}, x_{2,k_2} \ldots x_{n,k_n})$ informs the features of an object of interest. If a value, $x_{i,k_i} \in \mathbf{x}$, is not observed, its corresponding entry is empty. The most probable classification hypothesis c^* given the evidence $\mathbf{X}$ is computed as:

$$c^* = \arg \max_{c \in \{true, false\}} P(c|\mathbf{x}') = P(c) \prod_{i=1}^{n} P(x_{i,k_i}|c)\,.$$

Let it $\mathbf{D}$ be a supervised dataset with N instances defined on $\mathbf{X} \cup \{C\}$. If $\mathbf{D}$ is complete, it is possible to obtain a maximum likelihood estimate of naive Bayes parameters by calculating the relative marginal frequency for C and the conditional frequencies

for $X_i|C$. A dataset is complete when each entry of every instance has an assigned value. Otherwise, it is incomplete and contains missing values.

If a bacterium does not express a specific ribosomal protein, its PMF pattern is incomplete (Cuenod et al., 2021). If a feature is missing because the organism does not express the respective protein in the spectrum, data presents a class conditional missingness pattern (Lin and Haug, 2008). It is a missing not-at-random pattern (MNAR) and should not be ignored since it encodes evidence against/for a classification hypothesis.

The mainstream strategies for learning naive Bayes from incomplete data do not handle class conditional missing patterns. Data removal, which deletes all incomplete cases, could discard valuable information. Data imputation and Expectation-Maximization-like algorithms estimate missing values from the observed ones. Thus, they do not explore CPM as evidence. Furthermore, distinguishing between a missing at random (MAR) event and a MNAR one. An alternative is to proceed with feature stratification. In this technique, any missing entry of a feature X_i that presents an MNAR pattern is replaced by a constant value (*nob*) (Henry et al., 2013).

Stratification assumes categorical features or is encoded into a discretization algorithm. It copies with practices in the implementation of Bayesian classifiers, where discretization is a typical preprocessing step (Yang and Webb, 2009). Firstly because discretization avoids the necessity of estimating conditional densities for $P(X_i|y)$. It is a difficult task if you do not have a large sample or have limited knowledge about the underlying data distribution. Secondly, experimental evidence supports that discretization can improve naive Bayes performance (Mubaroq et al., 2019).

Tahan and Asadi (2018) also highlight that discretization can mitigate the impact of class imbalance on supervised learning. This is interesting because the class imbalance is a common condition in bioinformatics (Zhang et al., 2020). In binary classification, a dataset is considered class imbalanced if the number of cases labeled as c (the majority class) is significantly greater than the number of cases in $\overline{c}$ (the minority class). Let N_c and $N_{\overline{c}}$ be the number of cases in each class. Class imbalance is measured with the imbalanced ratio (IR = $\frac{N_c}{N_{\overline{c}}}$).

Let X_j be a feature with a continuous domain Ω_j. A supervised discretization algorithm aims at partitioning Ω_j in l intervals (or bins) so that the discretized X_j preserves its association with class labels. Navas-Palencia (2022) presents two mathematical programming-based procedures for solving supervised optimization problems given the following requirements: (a) missing values are binned in separate; (b) each bin has a minimal percentage of cases; and (c) the partitioning process should allow the computation of probabilistic divergence measures. Both of them try to maximize Jeffrey's divergence measure. The first algorithm employs backtracking constrained search (Hentenryck and Michel, 2013) and the second, mixed integer linear programming (Floudas, 1995).

3. MATERIAL AND METHODS

The effectiveness of stratification for dealing with CMP patterns was evaluated in two experiments during the development of binary classifiers for genus identification using MS fingerprints. The experiments were run on eight datasets extracted from the PUCHUY dataset (Silva et al., 2019). That database contains 14700 fingerprints distributed on 1133 genera. Each fingerprint has 58 entries corresponding to ribosomal proteins.

The datasets used in the tests were generated as follows. Firstly, the PUCHUY database was copied to a temporary file. Next, every instance belonging to the target genus was labeled as *true*, and the remainder were labeled as *false*.

The first and fifth columns of Table 1 list the datasets used in the tests and the respective target genus of each task. Table 1 also lists the number of selected features for each classification task (second and sixth columns) and the features/proteins that present a class CMP pattern (third and seventh columns). The fourth and eighth columns list the imbalance ratios of the datasets.

Table 1. Datasets and target genus of each test.

Genus	♯ feat.	CMP patterns	IR	Genus	♯ feat.	CMP patterns	IR
Bordetella	33	L7AE	22.9	Campylobacter	35	L7AE, L30	60.3
Corynebacterium	30	L7AE, S21	62.4	Escherichia	34	L7AE	15.2
Klebsiella	33	L7AE	31.0	Mycoplasma	41	L25, L30	65.2
Pseudomonas	31	L7AE	29.9	Salmonella	34	L7AE	21.4

Feature selection began by executing a procedure to identify class CMP patterns by checking which features, X_j, were not observed in the target class. These columns were removed from datasets used in tests with Gaussian naive Bayes but retained in datasets used with categorical classifiers. Next, the Mann-Whitney test was employed to verify whether there were differences between the readings of each class (Pérez et al., 2015). If a significant difference was found for an attribute X_j, it was preselected; otherwise, it was discarded. Finally, the relief algorithm (Kononenko, 1994), as implemented in CORELEARN R package[1] was run on all preselected features. Covariates whose relief index was higher than 0.02 were selected to form the final datasets. The threshold was experimentally optimized.

The first experiment consisted of training Gaussian and categorical naive Bayes classifiers (Reddy et al., 2022) on each dataset from Table 1. For Gaussian classifiers, missing data were imputed using the MICE algorithm (Buuren and Groothuis-Oudshoorn, 2011) implemented in Scikit-learn (Pedregosa et al., 2011). The training of categorical classifiers was performed on discretized/feature-stratified datasets. Data discretization was performed with the constrained programming and MILP-based supervised discretization procedures implemented in the OptBinning package Navas-Palencia (2022) (available at http://gnpalencia.org/optbinning). Feature stratification was applied after discretization.

That experiment was conducted with missing entry rates of 10%, 20%, and 30%. The classification performance under these conditions was estimated by calculating the average and standard deviation of balanced accuracy using 10-fold cross-validation. To compare the performance of categorical (feature-stratified datasets) and Gaussian (imputed datasets) classifiers, an ANOVA test with post hoc analysis was conducted. Table 2 lists each group considered in the ANOVA test.

The second experiment analyzed the impact of class imbalance on the performance of categorical naive Bayes trained on incomplete data treated with stratification.

[1] https://cran.r-project.org/web/packages/CORElearn/index.html

For that, a correlation analysis estimates the linear dependency between class imbalance and classifiers performance. Such analysis was repeated for each classifier and missing entries rate.

4. Results and discussion

Table 2 presents the experimental results. The first column enumerates the datasets. The next three pairs of columns list the mean balanced accuracy of each classifier for datasets with 20% and 30% of missing entries. The abbreviations GNB, NB-CP, and NB-MILP refer to Gaussian naive Bayes, naive Bayes with constrained programming-based discretization, and naive Bayes with mixed-integer LP-based discretization, respectively. The values in bold indicate the higher performances in each condition. The last two columns indicate whether ANOVA detected a significant difference among the groups of procedures described in the methodology (+) or not (-).

Table 2. Classification performance/missing rate and ANOVA.

Dataset	Balanced accuracy							
	GNB		NB-CP		NB-MILP		ANOVA	
	20%	30%	20%	30%	20%	30%	20%	30%
Bordetella	0.984	**0.985**	0.971	**0.985**	**0.989**	0.978	-	-
Campylobacter	0.992	0.985	0.991	0.987	**0.997**	**0.989**	-	-
Corynebacterium	0.991	0.992	0.994	0.972	**0.996**	**0.994**	-	+
Escherichia	0.944	0.935	**0.995**	**0.995**	**0.995**	**0.995**	+	+
Klebsiella	0.914	0.910	0.989	0.972	**0.990**	**0.981**	+	+
Mycoplasma	0.979	0.985	**0.996**	0.994	**0.996**	**0.996**	+	+
Pseudomonas	0.923	0.878	**0.994**	0.994	0.994	**0.995**	+	+
Salmonella	0.945	0.942	0.991	0.990	**0.995**	**0.995**	+	+

The results presented above indicate that categorical naive Bayes with MILP discretization and feature stratification achieved the highest accuracy in the majority of tests (7 out of 8). Gaussian naive Bayes with MICE imputation showed the best performance in only one test (Bordetella/30% of missing entries). The ANOVA test confirmed that the performance of categorical classifiers was consistently superior to that of Gaussian classifiers. Further evidence was provided by the *post hoc* analysis, which demonstrated that categorical classifiers outperformed Gaussian classifiers in most tests, regardless of the rate of missing entries.

The correlation analysis results are shown in Table 3. The first column indicates the proportion of missing entries in each dataset. Column IR $\times\Delta_1$ presents the correlation between the imbalanced ratio and the performance differences between NB-CP and GNB, while Column IR $\times\Delta_2$ does the same for GNB and NB-MILP. The negative values in these columns indicate a weak linear relationship between class imbalance and performance differences (Benesty et al., 2009). This suggests that the gains obtained through discretization and feature stratification are not significantly affected by class imbalance.

5. Conclusion

This work evaluated the effectiveness of feature stratification in exploring class-conditional missing patterns present in MALDI-TOF-MS fingerprint datasets used for

Table 3. Correlation analysis - IR $\times$ performance gain

Missingness rate	Correlation IR $\times\Delta_1$	Correlation IR $\times\Delta_2$
30%	-0.35	-0.48
50%	-0.39	-0.38

developing software for bacteria identification. The tests utilized the naive Bayes model for classification and employed two procedures to realize discretization and feature stratification. The ANOVA results showed that discretization/feature stratification significantly improved the predictive performance of classifiers in most of the tests. *Post hoc* analysis indicated that Gaussian naive Bayes fitted on imputed data never outperformed models trained on discretized datasets during the tests.

Furthermore, the correlation analysis demonstrated that the improvement in performance achieved by discretizing/stratifying the data is not strongly influenced by class imbalance. This finding is of practical interest, as MALDI-TOF-MS data used in bacteria identification often exhibit imbalanced class distributions.

Future research aims to investigate the relationship between the results obtained through the stratification of CMP patterns and factors such as class overlap and class separability.

6. Agradecimentos

CNPq, CAPES, Fundação Araucária, SETI.

References

Ashfaq, M., Da'na, D., and Al-Ghouti, M. (2022). Application of maldi-tof ms for identification of environmental bacteria: A review. *Journal of Envir. Manag.*, 305:114359.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). *Noise Reduction in Speech Processing*, chapter Pearson Correlation Coefficient, pages 1–4. Springer Berlin Heidelberg.

Buuren, S. and Groothuis-Oudshoorn, C. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67.

Cuenod, A., Foucault, F., Pfluger, V., and Egli, A. (2021). Factors associated with maldi-tof mass spectral quality of species identification in clinical routine diagnostics. *Frontiers in Cellular and Infection Microbiology*, 11.

de Souza, R., Ambrosini, A., and Passaglia, L. (2015). Plant growth-promoting bacteria as inoculants in agricultural soils. *Genetics and Molecular Biology*, 38:401 – 419.

Dong, Y. and Peng, J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2:222.

Floudas, C. (1995). *Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications*. Oxford Academic.

Haider, A., M. Ringer, Z. K., Mohacsi-Farkas, C., and Kocsis, T. (2023). The current level of maldi-tof ms applications in the detection of microorganisms: A short review of benefits and limitations. *Microbiology Research*, 14(1):80–89.

Henry, A. J., Hevelone, N. D., Lipsitz, S., and Nguyen, L. L. (2013). Comparative methods for handling missing data in large databases. *Journal of Vascular Surgery*, 58(5):1353–1359.e6.

Hentenryck, P. V. and Michel, L. (2013). The objective-cp optimization system. In *Proceedings of the 19th International Conference on Principles and Practice of Constraint Programming*, CP'13, page 8–29, Berlin, Heidelberg. Springer-Verlag.

Ke, X., Keenan, K., and Smith, V. (2022). Treatment of missing data in bayesian network structure learning: an application to linked biomedical and social survey data. *BMC Medical Research Methodology*, 22.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. In *Proceedings of the 7th European Conference on Machine Learning*, page 171–182, Berlin, Heidelberg. Springer-Verlag.

Lin, J.-H. and Haug, P. (2008). Exploiting missing clinical data in bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics*, 41(1):1–14.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, New York.

Mubaroq, T., Sugiharti, E., and Akhlis, I. (2019). Application of discretization and information gain on naive bayes to diagnose heart disease. *Journal of Advances in Information Systems and Technology*, 1(1):75–82.

Navas-Palencia, G. (2022). Optimal binning: mathematical programming formulation, arxiv:2001.08025.

Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, V. M. B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pérez, N., Guevara López, M., A.S., and Ramos, I. (2015). Improving the mann–whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. *Artificial Intelligence in Medicine*, 63(1):19–31.

Reddy, E., Gurrala, A., Hasitha, V., and Kumar, K. (2022). *Bayesian Reasoning and Gaussian Processes for Machine Learning Applications*, chapter Introduction to Naive Bayes and a Review on Its Subtypes with Applications, pages 1–14. Chapman and Hall/CRC eBooks.

Silva, R. R., Tomachewski, D., and ao, C. G. (Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial. BR512019002529-6, Nov. 2019). Banco de dados de massa molecular de proteÍnas ribossomais baseado em genomas bacterianos.

Tahan, M. and Asadi, S. (2018). Emdid: Evolutionary multi-objective discretization for imbalanced datasets. *Information Sciences*, 432:442–461.

Tomachewski., D., Galvão, C., de A. Campos Jr, Guimarães, A., da Rocha, J., and Etto, R. (2018). Ribopeaks: a web tool for bacterial classification through m/z data from ribosomal proteins. *Bioinformatics*, 34(17):3058–3060.

Weis, C., Jutzeler, C., and Borgwardt, K. (2020). Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. *Clinical Microbiology and Infection*, 26(10):1310–1317.

Yang, Y. and Webb, G. (2009). Discretization for naive-bayes learning: Managing discretization bias and variance. *Machine Learning*, 74:39–74.

Zhang, L., Ray, H., Priestley, J., and Tan, S. (2020). A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data. *Journal of Applied Statistics*, 47(3):568–581.

Zhang, R., Zhang, Y., Zhang, T., Xu, W., Wang, H., Zhang, S., Zhang, T., Zhou, W., and Shi, G. (2022). Establishing a maldi-tof-tof-ms method for rapid identification of three common gram-positive bacteria (bacillus cereus, listeria monocytogenes, and micrococcus luteus) associated with foodborne diseases. *Food Sci. and Tech.*, 42.