Fast Imbalanced Classification of Healthcare Data with Missing Values

Talayeh Razzaghi*, Oleg Roderick†, Ilya Safro* and Nick Marko†
*School of Computing
Clemson Univeristy, Clemson, SC 29634
Email: {trazzag, isafro}@clemson.edu
†Geisinger Health System
Danville, PA 17822
Email: {oroderick, nmarko}@geisinger.edu

***Abstract*—In medical domain, data features often contain missing values. This can create serious bias in the predictive modeling. Typical standard data mining methods often produce poor performance measures. In this paper, we propose a new method to simultaneously classify large datasets and reduce the effects of missing values. The proposed method is based on a multilevel framework of the cost-sensitive SVM and the expected maximization imputation method for missing values, which relies on iterated regression analyses. We compare classification results of multilevel SVM-based algorithms on public benchmark datasets with imbalanced classes and missing values as well as real data in health applications, and show that our multilevel SVM-based method produces fast, and more accurate and robust classification results.**

I. The role of predictive modeling in healthcare

Modern healthcare can be characterized as *evidence-driven* and *model-assisted* [1]. In an ideal situation, every decision in the clinical environment should be supported by a statistical model predicting risks and positive outcomes. This model may have a form of a simplified risk-assessment formula [2], or a sophisticated machine learning tool [3], [4]. In either case, it is based on a query of relevant clinical and operational history.

In practice, comprehensive medical information is stored in multiple databases, with different formats and rules of access. Due to considerations of patient privacy, and the proprietary nature of electronic medical records [5], the databases cannot be queried continuously. Every instance of data acquisition and integration is a separate effort that is cost-effective only when the resulting predictive model shows high quality. Thus, progress in evidence-driven healthcare depends on how well state-of-the art algorithms of machine learning are adapted to clinical data.

We note that classical computer science issues, such as scalability, or convergence rate are rarely a major issue for healthcare applications. Instead, an algorithm is ranked based on its ability to process raw medical data, with such problematic features as sparsity, missing entries, noise and imbalanced outputs. Because of the encounter nature of patient-provider interaction, medical data is inherently sparse: when a clinical encounter occurs, the number of and contents of labels attached to it vary widely [6]; outside of an encounter, the state of the patient is unknown. The outcomes of interest in classification problems are imbalanced, because, as a rule, healthcare analytics is motivated by rare events such as healthcare emergencies, severe chronic conditions, gaps and bottlenecks in access to care. The extent to which medical data is problematic may not be obvious from the perspective of a local healthcare provider (such as a single doctor); by definition they have access to all knowledge they ever use. We view this paper as a short, high-level primer on using advanced methods of machine learning to overcome the difficulties that emerge after multiple datasets are integrated for analysis and prediction.

This work was prompted by several projects completed with the Division of Applied Research and Clinical Informatics, Dept. of Data Science; Geisinger Health System. The routine activity of Data Science consists of medium-scope predictive projects on a combination of patient biometrics, pathology lab results, clinical encounter data, medical insurance data (available directly from Geisinger Health Plan) and externally assigned aggregate metrics for patients' general lifestyle risks, compliance with treatment regime, and loyalty to a particular provider.

For the first motivating example (Example 1), we use our 2014 feasibility study [7] of merging insurance information (6 aggregate features, based on the history of claims and payments) together with clinical encounter information (10-20 features chosen by hand from patient biometrics, medications and diagnostic codes). The goal of the initial study was to predict the financial risk for a particular patient (a common metric in insurance practice, derived as a ratio of individual expenses

and average expenses for a large demographic group). Furthermore, we wanted to see how addition of the clinical information changes the predictive power of the model, thus making a case for existence of high-risk patients that are invisible to claims-based analysis. We used a standard clustering technique, k-nearest neighbors with empirically selected weighting, to achieve the basic results.

For Example 2, we use our preliminary investigation of patients' response to public outreach [8], such as annual flu awareness campaigns. We included basic demographic and clinical information on patients targeted by 35 such campaigns into the model predicting whether a given patient is likely to respond to the reminder, or to choose not to get vaccinated, or use a different provider. Again, our core predictive model was standard: logistic regression with empirically selected weighting of training data.

We now pose a question: how much more effective would the predictive models be in each case with the use of an advanced machine learning algorithm developed with awareness of sparsity and class skewness (imbalance) in data?

II. Support Vector Machine Algorithms for Medical Data

Given a training set $\mathcal{J} = \{(x_i, y_i)\}_{i=1}^{l}$, that is a set of data points with known labels, where $(x_i, y_i) \in \mathbb{R}^{n+1}$, and l and n are the numbers of data points and features, respectively, and $y_i \in \{-1, 1\}$ denotes the class label for each data point i in $\mathcal{J}$. We denote by $\mathbf{C}^-$ and $\mathbf{C}^+$, the "majority" (points with $y_i = +1$) and "minority" (points with $y_i = -1$) classes respectively such that $\mathcal{J} = \mathbf{C}^+ \cup \mathbf{C}^-$.

A. Support Vector Machines

The support vector machine (SVM) solves the following max-margin problem:

$$\min \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \tag{1a}$$

$$\text{s.t.} \ \ y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \quad i = 1, \ldots, l \tag{1b}$$

$$\xi_i \geq 0 \quad i = 1, \ldots, l \tag{1c}$$

where the optimal margin is defined by the parameters w and b. The training data points x_i are mapped into a higher dimensional space through function $\phi : \mathbb{R}^n \to \mathbb{R}^m$ ($m \geq n$). The misclassified points are penalized using the term slack variables ξ_i ($i \in \{1, \ldots, l\}$) and the parameter $C > 0$ controls the magnitude of penalization. Hence, this formulation is called as *soft margin* SVM. The primal formulation is usually transformed to the Lagrangian dual problem using different algorithms. One of the most popular is the sequential minimal optimization (SMO) which is implemented in the LIBSVM tool [9], since it is fast and yields reliable convergence.

B. Weighted Support Vector Machines

A cost-sensitive extension of SVM, developed to cope with imbalanced data, is known as *weighted* SVM (WSVM) [10]. The main idea is to consider weighting scheme in learning such that the WSVM algorithm builds the decision hyperplane based on the relative contribution of data points in training. In contrast to the standard SVM, the penalization costs are different for the positive (C^+) and negative (C^-) classes:

$$\min \ \frac{1}{2}\|w\|^2 + C^+\sum_{\{i|y_i=+1\}}^{n_+}\xi_i + C^-\sum_{\{j|y_j=-1\}}^{n_-}\xi_j \tag{2a}$$

$$\text{s.t.} \ \ y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \quad i = 1, \ldots, l \tag{2b}$$

$$\xi_i \geq 0 \quad i = 1, \ldots, l \tag{2c}$$

The formulations (1) and (2) are solved through the Karush-Kuhn-Tucker conditions. The Gaussian kernel function (radial basis function, RBF) is used in the dual formulation of (W)SVMs since this kernel function usually results into superior performance for many classification problems [11, 12]. Parameter tuning is required to set optimal or near optimal C, C^+, C^-, and kernel function parameters (e.g. bandwidth parameter for RBF kernel function) to achieve good results for (W)SVM. This process becomes problematic and time-consuming particularly when the size of data is very large. Hence we aim to develop an efficient and effective classification method, called the Multilevel (W)SVM, that is scalable and works with imbalanced healthcare data.

C. Multilevel Support Vector Machines

The proposed algorithm belongs to the family of multilevel optimization strategies [13] whose goal is to approximate the system at multiple scales of coarseness and to obtain a final solution by combining the information from different scales. The multilevel framework for SVM [14] scales efficiently for large classification problems whose hierarchy of coarser representations is constructed based on the approximated k-nearest neighbors graphs (AkNN). This method consists of three main phases:

- **The coarsening phase.** A gradual coarsening of the training set is constructed using fast point selection method [15] in AkNN graph. However, we found that ensuring a uniform coverage of the points can lead to much better results than finding an independent set of points (nodes in AkNN) as was suggested in [15]. Thus, we extended the set of coarse points by setting a parameter for the

minimum number of points that in our experiments was set to 50% of the fine data points.

- **Supervised support vector initial learning.** After the hierarchy is created, the support vectors learning is performed at the coarsest level, where the number of data points is sufficiently small.
- **The uncoarsening phase.** Support vectors, and classifier are projected throughout the hierarchy from the coarsest to the finest levels. At each level, the solution to the current fine level is updated and optimized based on the solution of the previous coarse level. The locally optimal support vectors are obtained by gradual refinement of the projected support vectors from the coarse level.

For imbalanced data, the WSVM can easily be adopted as the base classifier for multilevel framework (ML-WSVM). The regular SVM does not perform well on imbalanced data because it tends to train models with respect to the majority class and technically ignores the minority class. However, the effect of imbalanced issue decreases while using multilevel framework since we prevent creating very small coarse sets for the minority class even if the majority class can still be coarsened.

Often, methods for imbalanced classification demonstrate poor performance on data with missing values (such as [16]) that is a frequent situation in healthcare data. Therefore, we apply imputation methods prior the classification model. Such imputation methods have been well studied in statistical analysis and machine learning domains [17–21]. Problems with missing data can be categorized into three types: data is completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR occurs while any feature of a data instance is missing completely random and is independent of the values of other features. Data is MAR , when the data instance with missing feature is dependent on the value of one or more of the instances other features. NMAR occurs when the data instance with missing feature is dependent on the value of the other missing features. Even though MCAR is more desirable, in many real-world problems, MAR occurs frequently in practice [17].

In the imputation methods, the goal is to substitute a missing value with a meaningful estimation [20]. This can be done either directly from the information on the dataset or by constructing a predictive model for this purpose. Standard methods for imputation are mean imputation [22], kNN imputation [23], Bayesian principal component analysis (BPCA) imputation [24], and the expectation maximization (EM) [25]. We apply the EM method which is one of the most successful imputation methods [26]. The EM method iteratively applies linear regression analysis and fits a new linear to the estimated data until a local optimum is achieved [25, 27]. In the regularized adaption of EM method, the conditional maximum likelihood estimation of regression parameters is replaced in the conventional EM algorithm [28].

D. Regularized Expectation-Maximization

In our preprocessing when the data contain many missing values, we apply the EM algorithm. It iteratively calculates the maximum-likelihood (ML) estimates of parameters by exploring the relationship between the complete data and the incomplete data (with missing features) [29]. In many cases, it has been demonstrated that the EM algorithm achieves reliable global convergence, economical storage. It is not computationally expensive, and can be easily implemented [30]. In EM we maximize the objective of the log-likelihood function

$$L(\Theta;\chi) = \sum_{i=1}^{n} logp(x_i|\Theta), \tag{3a}$$

where $\chi = \{x_i|i = 1, ..., n\}$ are the observations with independent distribuation $p(x)$ parametrized by Θ.

The regularized EM algorithm (REM) is developed to control the level of uncertainty associated to missing values [31]. The main idea is to regularize the likelihood function according to the mutual relationship between the observations and the missing data with little uncertainty and maximum information. Intuitively, it is desirable to select the missing data that has a high probabilistic association with the observations, which shows that there is little uncertainty on the missing data given the observations. It performs linear regression iteratively for the imputation of missing values. The REM algorithm optimizes the penalized likelihood as follows:

$$\tilde{L}(\Theta;\chi) = L(\Theta;\chi) + \gamma P(\chi, \Upsilon|\Theta), \tag{4a}$$

where P is the distribution function of the complete data given Θ. The trade-off between the degree of regularization of the solution and the likelihood function is controlled by the so-called regularization parameter that is represented by γ that [31]. In addition to reducing the uncertainty of missing data, the REM preserves the advantage of the standard EM method. This method is very efficient for over-complicated models.

E. Performance Measures

Classification algorithms are evaluated based on the performance measures, which are calculated from the confusion matrix (I). For binary classification problems, the performance measures are defined as accuracy (ACC), sensitivity (SN), specificity (SP), and G-mean, namely,

$$\text{SN} = \frac{TP}{TP + FN}, \quad \text{SP} = \frac{TN}{TN + FP} \tag{5}$$

TABLE I
CONFUSION MATRIX

	Positive class	Negative Class
Positive Class	True Positive (TP)	False Positive (FP)
Negative Class	False Negative (FN)	True Negative (TN)

$$\text{G-mean} = \sqrt{\text{SP} * \text{SN}} \tag{6}$$

$$\text{ACC} = \frac{TP + TN}{FP + TN + TP + FN}. \tag{7}$$

III. COMPUTATIONAL RESULTS

We evaluate the proposed classification framework on academic (UCI [32], and the cod-rna dataset [33]), and real-life binary classification benchmarks [7], [8]. Coarsest and refinement (W)SVM models are solved using LIBSVM-3.18 [9], and the FLANN library [34] is used to create the AkNN graphs. Multilevel frameworks, data processing and further scripting are implemented in MATLAB 2012a. The C4.5, Naive Bayes (NB), Logistic Regression (LR), and 5-Nearest Neighbor (5NN) are implemented using WEKA interfaced with MATLAB. A typical 10-fold cross validation setup is used. We create missing values on the academic data training sets by discarding the features randomly. The misclassification penalty or weights are selected as inversely proportional to the size of each class in our implementation. As a preprocessing step, the whole data is normalized before classification. The nested uniform design (UD) is performed on the training data as the model selection for (W)SVM [35]. The UD methodology is very successful for model selection in supervised learning [36]. The close-to-optimal parameter set is achieved in an iterative nested process [35]. The optimal parameter set is selected based on *G-mean* maximization, since data might be imbalanced. A 9- and 5-point run design is performed for the first and second stages of the nested UD due to its superiority for the UCI data [35], and the performance measures such as sensitivity, specificity, G-mean and accuracy are calculated on the testing data.

TABLE II
ACADEMIC DATA SETS.

Dataset	r_{imb}	n_f	$\lvert\mathcal{J}\rvert$	$\lvert\mathbf{C}^+\rvert$	$\lvert\mathbf{C}^-\rvert$
Twonorm	0.50	20	7400	3703	3697
Letter26	0.96	16	20000	734	19266
Ringnorm	0.50	20	7400	3664	3736
Cod-rna	0.67	8	59535	19845	39690
Clean (Musk)	0.85	166	6598	1017	5581
Advertisement	0.86	1558	3279	459	2820
Nursery	0.67	8	12960	4320	8640
Hypothyroid	0.94	21	3919	240	3679
Buzz	0.80	77	140707	27775	112932

A. Academic data sets

We compared popular methods with the proposed ML(W)SVM to classify imperfect data. Table III shows the comparative results of MLSVM, MLWSVM, SVM, WSVM, Naive Bayes, C4.5, LR, and 5NN algorithms for academic data sets. These methods are examined for different missing value ratios selected as 5%, 10%, 20%, and 40%. We implemented the REM method for missing data imputation [25]. The highest values are shown in boldface among all methods for their related missing value levels. It is clear from the accumulation of boldface results, MLWSVM and WSVM perform better than the other methods in general for all missing vaule ratios. In fact, MLWSVM and WSVM results into higher G-mean values in 19 out of 36 dataset/r_{mv} combinations followed by MLSVM and SVM with 13 out of 36. Moreover, the ML(W)SVM techniques achieve faster computational time compared to the standard (W)SVM (Table IV).

TABLE IV
COMPUTATIONAL TIME (SEC.)

	MLSVM	SVM	MLWSVM	WSVM
Twonorm	6	29	6	29
Letter	37	145	39	146
Ringnorm	5	26	5	27
Cod-rna	300	1865	315	1891
Clean	25	103	23	90
Advertiesment	99	228	101	232
Nursery	26	188	32	193
Hypothyroid	2	3	2	3
Buzz	3915	26963	4705	27732

B. Healthcare data sets

We present the results of comparison of classification algorithms on the real-life healthcare data sets. Table V demonstrates the results on Example 1 (see Section I), a classification task of assigning a patient in a correct group by financial risk, which are ordered in ascending manner from group 1 with the lowest level of risk, to group 5 with the highest level of risk.

The motivation behind the original study was to determine how much integration of the medical and financial data changes the outcomes of clustering and classification operations based on financial data alone. For that purpose, modest precision was sufficient; we used a logistic (linear) regression (LR) approach (implemented as *mnrfit* in MATLAB). We are comparing the accuracy of it with the best results obtained by ML(W)SVM. The strategy "one-against-all" is used for multi-class classification. This strategy performs training a classifier per class with the data points of that class as positive class and the rest of the data points are trained as negative class.

TABLE III
COMPARATIVE G-MEAN RESULTS FOR ML(W)SVM AGAINST THE REGULAR SVM, WSVM, NB, C4.5, 5NN, AND LR ON ACADEMIC DATASETS FOR DIFFERENT FRACTIONS OF MISSING VALUES (r_{mv}).

Dataset	r_{mv}	MLSVM	MLWSVM	SVM	WSVM	C4.5	5NN	NB	LR
Twonorm	5%	**0.98**	**0.98**	**0.98**	**0.98**	0.86	0.97	**0.98**	**0.98**
	10%	**0.98**	**0.98**	0.97	0.97	0.87	0.97	0.97	0.97
	20%	**0.98**	**0.98**	**0.98**	**0.98**	0.88	0.97	0.97	**0.98**
	40%	0.97	0.97	0.97	0.97	0.89	0.97	**0.98**	**0.98**
Letter	5%	0.97	**1.00**	0.99	0.99	0.97	0.98	0.86	0.81
	10%	0.98	**1.00**	0.98	0.99	0.98	0.98	0.86	0.80
	20%	**1.00**	**1.00**	0.99	0.99	0.97	0.98	0.87	0.80
	40%	0.95	0.97	0.96	**0.99**	0.97	0.98	0.88	0.83
Ringorm	5%	0.97	0.98	0.97	0.98	0.91	0.61	**0.99**	0.76
	10%	0.98	0.98	**0.99**	**0.99**	0.91	0.62	0.98	0.76
	20%	**0.98**	**0.98**	0.97	**0.98**	0.91	0.62	**0.98**	0.76
	40%	**0.98**	**0.98**	0.97	**0.98**	0.91	0.62	**0.98**	0.76
Cod-rna	5%	0.95	**0.96**	**0.96**	**0.96**	0.95	0.92	0.66	0.93
	10%	0.95	**0.96**	0.95	0.96	0.95	0.91	0.66	0.92
	20%	0.95	**0.96**	0.95	0.95	0.94	0.91	0.67	0.92
	40%	**0.95**	**0.95**	**0.95**	**0.95**	0.93	0.90	0.68	0.91
Clean	5%	**1.00**	0.99	0.98	**1.00**	0.83	0.92	0.79	0.89
	10%	0.99	**1.00**	0.99	**1.00**	0.83	0.91	0.79	0.89
	20%	**1.00**	**1.00**	**1.00**	**1.00**	0.83	0.91	0.79	0.89
	40%	**1.00**	**1.00**	**1.00**	**1.00**	0.82	0.92	0.79	0.89
Advertisement	5%	0.87	0.87	0.87	0.87	**0.92**	0.81	0.60	0.82
	10%	**0.87**	**0.87**	0.86	0.86	0.86	0.85	0.62	0.82
	20%	0.83	0.85	0.83	0.85	**0.89**	0.83	0.61	0.83
	40%	0.84	0.86	0.87	0.81	**0.91**	0.85	0.62	0.82
Nursery	5%	0.99	0.99	**1.00**	**1.00**	**1.00**	**1.00**	0.00	**1.00**
	10%	0.99	0.99	**1.00**	**1.00**	**1.00**	**1.00**	0.00	**1.00**
	20%	0.96	0.96	**1.00**	**1.00**	**1.00**	**1.00**	0.00	1.00
	40%	0.92	0.92	**1.00**	**1.00**	**1.00**	0.99	0.46	**1.00**
Hypothyroid	5%	0.83	0.87	0.81	0.87	0.96	0.76	**0.97**	0.88
	10%	0.85	0.86	0.78	0.86	**0.96**	0.76	**0.96**	0.89
	20%	0.84	0.86	0.72	0.86	0.96	0.75	**0.97**	0.90
	40%	0.86	0.88	0.84	0.88	0.96	0.76	**0.97**	0.89
Buzz	5%	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	0.93	0.89	**0.94**
	10%	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	0.93	0.89	**0.94**
	20%	0.92	**0.94**	0.93	**0.94**	**0.94**	0.93	0.88	0.93
	40%	0.93	0.93	0.93	0.93	**0.94**	**0.94**	0.86	**0.94**

TABLE V
ACCURACY OF FINANCIAL RISK PROBLEM WITH FIVE RISK CLASSES (EXAMPLE 1)

Class	1	2	3	4	5
LR	0.58	0.54	0.53	0.51	0.59
MLSVM	0.73	0.50	0.44	0.50	0.71

To interpret the results, we note that correct identification of intermediate risk categories is a very difficult problem in medical informatics. To our knowledge, there is no good definition of "average health", either evidence-driven or philosophical, that would help an expert to identify such patient features that do not indicate an acute crisis, or an almost certain safety from crisis. Accordingly, it is not surprising that neither approach does well on the risk categories 2..4; there is also not a lot of motivation to improve the model there. On the other hand, it is important to identify and predict the very low-risk patients (knowing that status ahead of time allows resource re-allocation leading to savings and improved service for everyone) and the very high-risk patients (so that clinical and financial resources could be prepared for the forthcoming crisis).

Accordingly, it is important that the use of an advanced method of machine learning changes the quality of prediction from almost worthless ('toss a coin') to workable (accuracy of 0.7).

In Table VI, we compare results for the widely used basic approach and ML(W)SVM prediction for Example 2 (see Section I), a study of patient's response to hospital flu outreach. In this problem, the goal is to find a binary classifier that will predict whether the patient will get vaccinated after reminder, or not (this includes using a different provider for vaccination). In the preliminary study, we used adaptive linear regression model (LASSO for adaptive selection of features, logistic regression on actual prediction).

Response to outreach is not a crucial life-or-death issue, we are performing this study to see if predictive modeling can assist with resource allocation (which patients to contact, how much medical personnel effort

TABLE VI
COMPARISON OF MULTILEVEL WSVM AGAINST MULTILEVEL SVM AND ADAPTIVE LOGISTIC REGRESSION (LR). IMPROVED RESULTS ARE IN BOLD.

	G-mean	SN	SP	ACC
Adaptive LR	0.7516	0.8903	0.6345	0.7619
MLSVM	**0.8012**	**0.9750**	**0.6583**	**0.8496**
MLWSVM	**0.8016**	**0.9739**	**0.6598**	**0.8495**

to dedicate to outreach and then vaccination). Arguably, accuracy is more important than specificity here. Even the basic results (using linear regression) were met with approval the CPSL (Care Patient Service Line: a division responsible for coordinating efforts of local, small-scale healthcare providers operating under Geisinger). SVM methods (almost 10 percent improvement) provide additional justification for the use of machine learning on merged data to assist planning in clinical practice.

IV. CONCLUSION

Large-scale data, missing or imperfect features, skewness distribution of classes are common challenges in pattern recognition of many healthcare problems. We have successfully extended a powerful machine learning technique, support vector machines, to the *scalable multilevel framework* of cost-sensitive learning SVM to deal with imbalanced classification problems. Our multilevel framework substantially improves the computational time without losing the quality of classifiers for large-scale datasets. We have shown that MLWSVM produces superior results than MLSVM and the regular SVM methods in most cases. This work can be extended to tackle other classification problems with large-scale imbalanced data (combined from different sources) with missing features in healthcare and engineering applications.

From the perspective of evidence-driven healthcare, our work shows that application of cutting edge machine learning techniques (in this case, fast multilevel classifiers) makes enough of a difference to justify the additional development effort for typical examples from clinical practice. While the improvements in precision and specificity we show in this study are both under 10% and are modest in general perspective, the result in healthcare is significant.

To our knowledge, such complex combined behavioral/operational phenomena as inference of financial risk from medical history (Example 1), or prediction of effectiveness of public outreach (Example 2), don't have a satisfactory casual explanation. The classical (1990s) clinical practice offered two equally unsatisfactory options: not having a capability for prediction at all, or relying on very basic statistical techniques (based on a single data source, with very high rate of false-positive classification outcomes). The existing mature models (such as actuarial projections of financial risk) do not benefit from integration of data from multiple sources, and may, in fact, turn out to be ineffective outside of their scope in patient population and metrics of interest (as we have shown in [7]). Thus, in the modern clinical practice we have to rely on newly developed machine learning tools, tuned on data from multiple sources. Thus, our work can also be extended to handle other classification problems on massive, multi-format medical data.

REFERENCES

[1] S. Foldy, "National public health informatics, united states," in *Public Health Informatics and Information Systems*. Springer, 2014, pp. 573–601.

[2] L. R. Haas, P. Y. Takahashi, N. D. Shah, R. J. Stroebel, M. E. Bernard, D. M. Finnie, and J. M. Naessens, "Risk-stratification methods for identifying patients for care coordination." *The American journal of managed care*, vol. 19, no. 9, pp. 725–732, 2013.

[3] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz, "A machine learning approach to predicting blood glucose levels for diabetes management," *Modern Artificial Intelligence for Health Analytics. Papers from the AAAI-14*, 2014.

[4] L. K. Woolery and J. Grzymala-Busse, "Machine learning for an expert system to predict preterm birth risk," *Journal of the American Medical Informatics Association*, vol. 1, no. 6, pp. 439–446, 1994.

[5] E. Larson, T. Bratts, J. Zwanziger, and P. Stone, "A survey of irb process in 68 us hospitals," *Journal of Nursing Scholarship*, vol. 36, no. 3, pp. 260–264, 2004.

[6] C. Snomed, "Systematized nomenclature of medicine-clinical terms," *International Health Terminology Standards Development Organisation*, 2011.

[7] O. Roderick, "Why risk models and access to large data should work together: tangible advantages to merging clinical and claims data," *Technical report, DARCI: Data Science, Geisinger Health System*, 2014.

[8] ——, "Predictive model for response to medical outreach," *Technical report, DARCI: Data Science, Geisinger Health System*, 2015.

[9] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[10] K. Veropoulos, C. Campbell, N. Cristianini *et al.*, "Controlling the sensitivity of support vector ma-

chines," in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 1999, 1999, pp. 55–60.

[11] F. E. Tay and L. Cao, "Application of support vector machines in financial time series forecasting," *Omega: The International Journal of Management Science*, vol. 29, no. 4, pp. 309–317, 2001.

[12] P. Xanthopoulos and T. Razzaghi, "A weighted support vector machine method for control chart pattern recognition," *Computers & Industrial Engineering*, vol. 70, pp. 134–149, 2014.

[13] A. Brandt and D. Ron, "Chapter 1 : Multigrid solvers and multilevel optimization strategies," in *Multilevel Optimization and VLSICAD*, J. Cong and J. R. Shinnerl, Eds. Kluwer, 2003.

[14] T. Razzaghi and I. Safro, "Fast multilevel support vector machines," *Preprint ArXiv*, 2014.

[15] S. Sakellaridi, H. ren Fang, and Y. Saad, "Graph-based multilevel dimensionality reduction with applications to eigenfaces and latent semantic indexing," in *Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference on*, Dec 2008, pp. 194–200.

[16] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognition*, vol. 41, no. 12, pp. 3692–3705, 2008.

[17] M. Ghannad-Rezaie, H. Soltanian-Zadeh, H. Ying, and M. Dong, "Selection–fusion approach for classification of datasets with missing values," *Pattern recognition*, vol. 43, no. 6, pp. 2340–2350, 2010.

[18] R. J. Little and D. B. Rubin, "Statistical analysis with missing data," 2002.

[19] J. L. Schafer, *Analysis of incomplete multivariate data*. CRC press, 2010.

[20] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010.

[21] I. A. Gheyas and L. S. Smith, "A neural network-based framework for the reconstruction of incomplete data sets," *Neurocomputing*, vol. 73, no. 16, pp. 3039–3065, 2010.

[22] A. R. T. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons, "Review: a gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006.

[23] G. E. Batista and M. C. Monard, "A study of k-nearest neighbour as an imputation method." *HIS*, vol. 87, pp. 251–260, 2002.

[24] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii, "A bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.

[25] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[26] B. C. Huang and A. Salleb-Aouissi, "Maximum entropy density estimation with incomplete presence-only data," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 240–247.

[27] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an em approach," in *Advances in Neural Information Processing Systems 6*. Citeseer, 1994.

[28] L. Nanni, A. Lumini, and S. Brahnam, "A classifier ensemble approach for the missing feature problem," *Artificial intelligence in medicine*, vol. 55, no. 1, pp. 37–50, 2012.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[30] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM review*, vol. 26, no. 2, pp. 195–239, 1984.

[31] H. Li, K. Zhang, and T. Jiang, "The regularized em algorithm," in *Proceedings of the national conference on artificial intelligence*, vol. 20, no. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005, p. 807.

[32] A. Frank and A. Asuncion, "UCI machine learning repository," *[http://archive. ics. uci.edu/ml]. irvine, ca: University of california," School of Information and Computer Science"*, vol. vol. 213, 2010.

[33] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[34] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *International Conference on Computer Vision Theory and Application VISSAPP'09)*. INSTICC Press, 2009, pp. 331–340.

[35] C. Huang, Y. Lee, D. Lin, and S. Huang, "Model selection for support vector machines via uniform design," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 335–346, 2007.

[36] O. L. Mangasarian and E. W. Wild, "Privacy-preserving classification of horizontally partitioned

data via random kernels.” in *DMIN*, 2008, pp. 473–479.