

A Discriminative Feature Space for Detecting and Recognizing Pathologies of the Vertebral Column

Damian Mingle*

WPC Healthcare, 1802 Williamson Court I, Brentwood, USA

Abstract

Each year it has become more and more difficult for healthcare providers to determine if a patient has a pathology related to the vertebral column. There is great potential to become more efficient and effective in terms of quality of care provided to patients through the use of automated systems. However, in many cases automated systems can allow for misclassification and force providers to have to review more causes than necessary. In this study, we analyzed methods to increase the True Positives and lower the False Positives while comparing them against state-of-the-art techniques in the biomedical community. We found that by applying the studied techniques of a data-driven model, the benefits to healthcare providers are significant and align with the methodologies and techniques utilized in the current research community.

Keywords: Vertebral column; Feature engineering; Probabilistic modeling; Pattern recognition

Introduction

Over the years there has been an increase in machine learning (ML) techniques, such as Random Forest (RF), Boosting (ADA), Logistic (GLM), Decision Trees (RPART), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) applied to many medical fields. A significant reason this has become the case is the capacity for human beings to act as diagnostic tools over time. Stress, fatigue, inefficiencies, and lack of knowledge all become barriers to high-quality outcomes.

There have been studies regarding applications of data mining in different fields, namely: biochemistry, genetics, oncology, neurology and EEG analysis. However, literature suggests that there are few comparisons of machine learning algorithms and techniques in medical and biological areas. Of these ML algorithms, the most common approach to develop nonparametric and nonlinear classifications is based on ANNs.

In general, the numerous methods of machine learning that have been applied can be grouped into two sets: knowledge-driven models and data-driven models. The parameters of the knowledge-driven models are estimated based on the expert knowledge of detecting and recognizing pathologies of the vertebral column. On the other hand, the parameters of data-driven models are estimated based on quantitative measures of associations between evidential features within the data. The classification models used in pathologies of the vertebral column have been SVM.

Studies have shown that ML algorithms are more accurate than statistical techniques, especially when the feature space is more complex or the input datasets are expected to have different statistical distributions [1]. These algorithms have the potential to identify and model the complex non-linear relationships between the features of the biomedical data set collected by Dr. da Mota, namely: pelvic incidence (PI), pelvic tilt (PT), lumbar lordosis angle (LLA), sacral slope (SS), pelvic radius (PR), and grade of spondylolisthesis (GOS).

These methods can handle a large number of evidential features that may be important in detecting abnormalities in the vertebral column. However, increasing the number of input evidential features may lead to increased complexity and larger numbers of model parameters, and in turn the model becomes susceptible to over fitting due to the curse of dimensionality.

This work aims to present medical decision support for those healthcare providers who are working to diagnosis pathologies of the vertebral column. This framework is comprised of three subsystems: feature engineering, feature selection, and model selection.

Pathologies of the vertebral column

Vertebrae, intervertebral discs, nerves, muscles, medulla, and joints make up the vertebral column. The essential functions of the vertebral column are as follows: (i) human body support (ii) protection of the nervous roots and medulla spine; and (iii) making the body's movement possible [2].

The structure of the intervertebral disc can be injured due to small or several small traumas in the column. Various pathologies can cause intense pain, such as disc hernias and spondylolisthesis. Backaches can be the results of complications that are caused within this complex system. We briefly characterize the biomechanical attributes that represent each patient in the data set.

Patient characteristics: Dr. Henrique da Mota collected data on 310 patients from sagittal panoramic radiographies of the spine while at the Centre Medico-Chirurgical de Readaptation des Massues placed in Lyon, France [3]. 100 patients were volunteers that had no pathology in their spines (labeled as 'Normal'). The remainder of patients had disc hernia (60 patients) or spondylolisthesis (150 patients).

Decision support for orthopedists is automated using ML algorithms and techniques of real clinical cases that utilize the above biomechanical attributes. Following, we compare many ML models evaluated through this study.

Problem statement and standard solutions

Classification refers to the problem of categorizing observations

*Corresponding author: Damian Mingle, Chief Data Scientist, WPC Healthcare, 1802 Williamson Court I, Brentwood, USA, Tel: 615-364-9660; E-mail: dmingle@wpchealthcare.com

Received June 30, 2015; Accepted August 19, 2015; Published September 15, 2015

Citation: Mingle D (2015) A Discriminative Feature Space for Detecting and Recognizing Pathologies of the Vertebral Column. Biomedical Data Mining 4: 114. doi:10.4172/2090-4924.1000114

Copyright: © 2015 Mingle D. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

into classes. Predictive modeling uses samples of data for which the class is known to generate a model for classifying new observations. We are only interested in two possible outcomes: ‘Normal’ and ‘Abnormal’. Complex datasets make it difficult not to misclassify some observations. However, our goal was to minimize those errors using the receiver operating characteristic (ROC) curve.

Literature suggests using an ordinal data approach for detecting reject regions in combinations with SVM. In addition, selecting the misclassification costs as follows: C_{low} cost when classifying a class as reject and assign C_{high} cost when misclassifying.

Therefore, $Reject = C_{low} / C_{high} = wr$ is the cost of rejecting (normalized by the cost of erring). The method accounts to account for the rejections rate and the misclassification rate [2].

Description of the data

It is useful to understand the basic features of the data in our study. Simple summaries about the sample and the measures, together with graphical analysis, form a solid basis for our quantitative analysis of the vertebral column dataset. We conducted univariate analysis which identifies the distribution, central tendency, and dispersion of the data.

The distribution table include the 1st and 3rd quartile, indicating 25% of the values that the observations demonstrate are less than or greater than the values listed (Table 1).

Distributions: Distribution of Biomechanical Features in class is specified in Figure 1.

Correlation: A correlation analysis provides insights into the independence of the numeric input variables. Modeling often assumes independence, and better models will result when using independent input variables. Below is a table of the correlations between each of the variables (Table 2).

We made use of a Hierarchical dendrogram to provide visual clues to the degree of closeness between variables [4]. The hierarchical correlation dendrogram produced here presents a view of the variables

of the dataset showing their relationships. The purpose is to efficiently locate groupings of variables that are highly correlated. The length of the lines in the dendrogram provides a visual indication of the degree of correlation. For example, shorter lines indicate more tightly correlated variables (Figure 2).

The feature engineering and data replication method

We developed a method which we termed Feature Bayes. This method makes use of a probabilistic model from synthetic data creation. Additionally, the data has been feature engineered and further refined through automated feature selection. In order to maximize prediction accuracy we generated 54 additional features. We define a row vector as $A = [a_1 \ a_2 \ \dots \ a_6]$ using the original six features from the vertebral column dataset. N is defined as the number of terms.

The features were constructed as follows:

‘Trim mean 80%’ calculates the mean taken by excluding a percentage of data points from the top and bottom tails of a vector as such

$$\bar{x} = \frac{\sum_{ij} A}{N} \tag{1}$$

Information theory, ‘Entropy’, is the expected value of the information contained in each message received [5] and is generally constructed as

$$\log_2 \sum_{n=a_i}^{a_o} n \tag{2}$$

‘Range’ is known as the area of variation between upper and lower limits and is generally defined as

$$A_{max} - A_{min} \tag{3}$$

We developed ‘Standard Deviation of A’ as a quantity calculated to indicate the extent of Deviation for a group as a whole,

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \tag{4}$$

	Pelvic_Incidence	Pelvic_Tilt	Lumbar_Lordosis_Angle	Sacral_Slope	Pelvic_Radius	Degree_Spondylolisthesis
Minimum	26.15	-6.555	14	13.37	70.08	-11.058
1st quarter	45.7	10.759	36.64	33.11	110.66	1.474
Median	59.6	16.481	49.78	42.65	118.15	10.432
Mean	60.96	17.916	52.28	43.04	117.54	27.525
3rd quarter	74.01	21.936	63.31	52.55	125.16	42.81
Maximum	129.83	49.432	125.74	121.43	157.85	418.543
	Class					
Abnormal	145					
Normal	72					

Table 1: Descriptive statistics of sample data

Correlation summary using the 'Pearson' covariance						
	pelvic_radius	pelvic_tilt	degree_spondylolisthesis	lumbar_lordosis_angle	sacral_slope	pelvic_incidence
pelvic_radius	1	0.01917945	-0.04701219	-0.04345604	-0.34769211	-0.25869222
pelvic_tilt	0.01917945	1	0.37008759	0.45104586	0.04615349	0.6307171
degree_spondylolisthesis	-0.04701219	0.37008759	1	0.50847068	0.55060557	0.6478843
lumbar_lordosis_angle	-0.04345604	0.45104586	0.50847068	1	0.53161132	0.6812879
sacral_slope	-0.34769211	0.04615349	0.55060557	0.53161132	1	0.8042957
pelvic_incidence	-0.25869222	0.63071714	0.64788429	0.68128788	0.80429566	1

*Note that only correlations between numeric variables are reported

Table 2: Pearson correlation matrix (Sample)

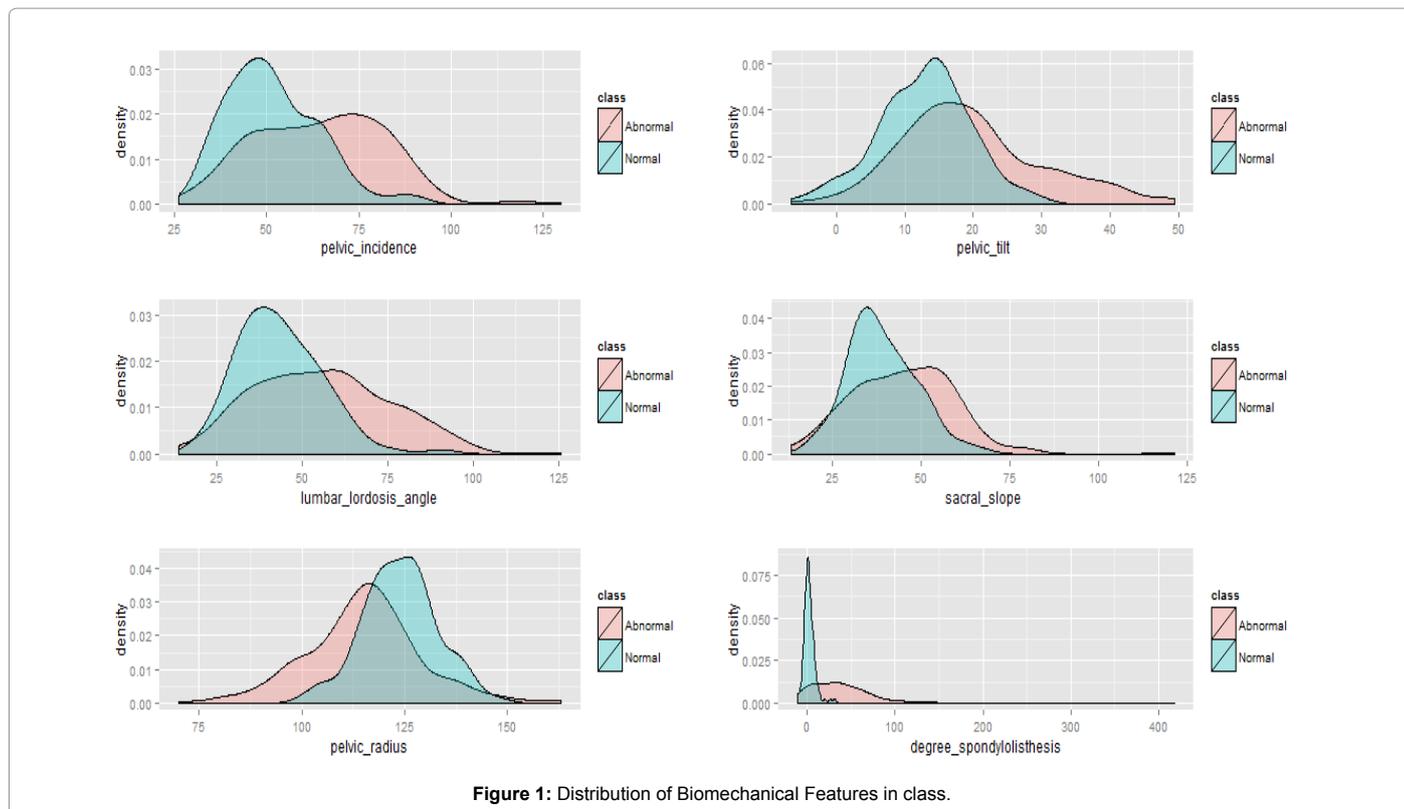


Figure 1: Distribution of Biomechanical Features in class.

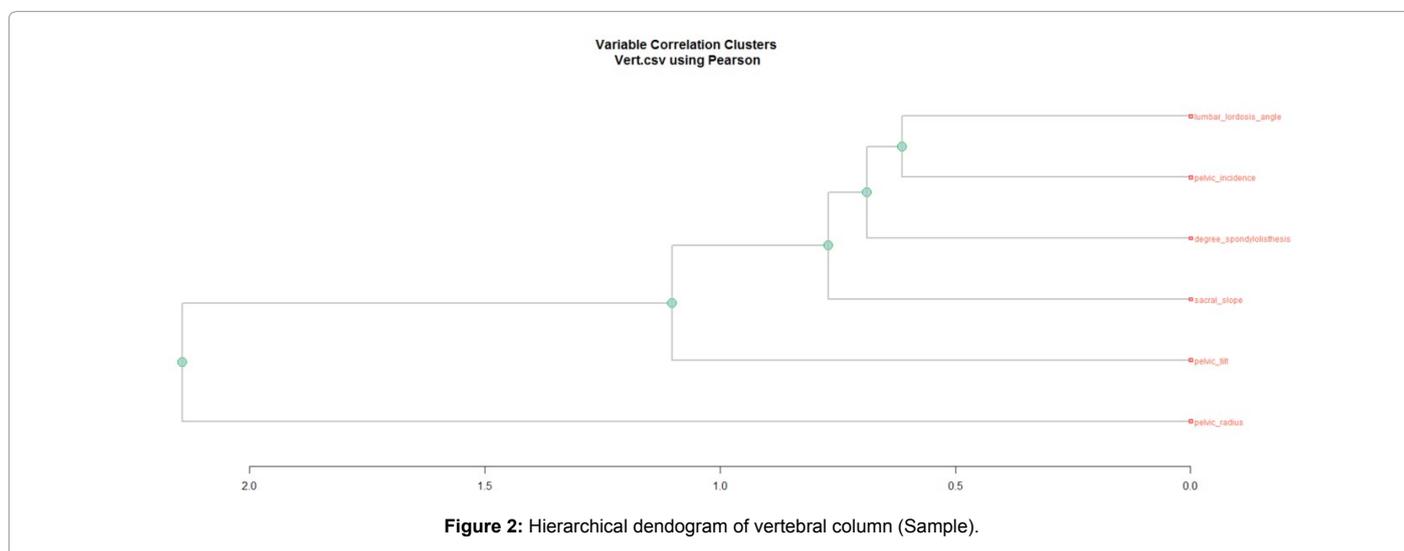


Figure 2: Hierarchical dendrogram of vertebral column (Sample).

‘Cosine of A ’ was generated to capture the trigonometric function that is equal to the proportion of the adjacent side to an acute angle of the hypotenuse,

$$\cos A \quad (5)$$

‘Tangent of A ’ was generated to capture the trigonometric function equal to the proportion of the opposite side over the adjacent side in a right triangle,

$$\tan A \quad (6)$$

‘Sine of A ’ was generated to capture the trigonometric function that is equal to the relationship of the opposite side of a given angle to the hypotenuse,

$$\sin A \quad (7)$$

‘25th Percentile of A ’ is the value of vector A such that 25% of the relevant population is below that value,

$$25th \text{ Percentile} = \left\lceil \frac{25}{100} * N \right\rceil \quad (8)$$

'20th Percentile of A' is the value of vector A such that 20% of the relevant population is below that value,

$$20th\ Percentile = \left\lceil \frac{20}{100} * N \right\rceil \quad (9)$$

'75th Percentile of A' is the value of vector A such that 75% of the relevant population is below that value

$$75th\ Percentile = \left\lceil \frac{75}{100} * N \right\rceil \quad (10)$$

'80th Percentile of A' is the value of vector A such that 80% of the relevant population is below that value,

$$80th\ Percentile = \left\lceil \frac{80}{100} * N \right\rceil \quad (11)$$

'Pelvic Incidence Squared' was used to change the pelvic incidence from a single dimension into an area. Many physical quantities are integrals of some other quantity,

$$a_1^2 \quad (12)$$

For each element of the row vector A we performed a square root calculation that yields a definite quantity when multiplied by itself,

$$\sqrt{a_{i,j}} \quad (13)$$

For each element of the row vector A we created a 'Natural Log of ai,j', more specifically a logarithm to the base of e

$$\ln a_{ij} \quad (14)$$

'Sum of pelvic incidence and pelvic tilt',

$$\sum_{n=a_1}^{a_2} n \quad (15)$$

For each element of the row vector A we created a 'Cubed' value of ai,j,

'Difference of pelvic incidence and pelvic tilt',

$$a_{i,j}^3 \quad (16)$$

'Difference of pelvic incidence and pelvic tilt',

$$a_1 - a_2 \quad (17)$$

'Product of pelvic incidence and pelvic tilt',

$$\prod_{n=a_1}^{a_2} n \quad (18)$$

'Sum of pelvic tilt and lumbar lordosis angle',

$$\sum_{n=a_2}^{a_3} n \quad (19)$$

'Sum of lumbar lordosis angle and sacral slope',

$$\sum_{n=a_3}^{a_4} n \quad (20)$$

'Sum of pelvic radius and degree spondylolisthesis',

$$\sum_{n=a_4}^{a_5} n \quad (21)$$

'Difference of pelvic tilt and lumbar lordosis angle',

$$a_2 - a_3 \quad (22)$$

'Difference of lumbar lordosis angle and sacral slope'

$$a_3 - a_4 \quad (23)$$

'Difference of sacral slope and pelvic radius

$$a_4 - a_5$$

Difference of pelvic radius and degree spondylolisthesis',

$$a_5 - a_6 \quad (25)$$

'Quotient of pelvic tilt and pelvic incidence',

$$\frac{a_2}{a_1} \quad (26)$$

'Quotient of lumbar lordosis angle and pelvic tilt',

$$\frac{a_3}{a_2} \quad (27)$$

'Quotient of sacral slope and lumbar lordosis angle',

$$\frac{a_4}{a_3} \quad (28)$$

'Quotient of pelvic radius and sacral slope',

$$\frac{a_5}{a_4} \quad (29)$$

'Quotient of degree spondylolisthesis and pelvic radius',

$$\sum_{n=a_4}^{a_5} n \quad (30)$$

'Sum of elements A',

$$\sum_{n=a_4}^{a_5} n \quad (31)$$

'Average of A elements',

$$\bar{x}_A \quad \dots\dots\dots(32)$$

'Median of A elements',

$$Median = \frac{\left(\frac{n}{2}\right)^{th} term + \left(\frac{n}{2} + 1\right)^{th} term}{2} \quad (33)$$

'Euler's number raised to the power of ai,j',

$$e^{a_{ij}} \quad (34)$$

Patient data generated with oversampling

The category 'Normal' was significantly underrepresented in the dataset. We employed the Synthetic minority oversampling technique (SMOTE) [6]. We chose the class value 'Normal' to work with using five nearest neighbors to construct an additional 100 instances.

Algorithm SMOTE (T,N,k)

Input: Number of minority class samples T; Amount of SMOTE N%; Number of nearest neighbors k

Output: (N/100) * T synthetic minority class samples

1. (* If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd*)
2. **If** N<100
3. **then** Randomize the T minority class samples
4. T=(N/100) * T
5. N=100
6. **end if**
7. N=(int)(N/100) (*The amount of SMOTE is assumed to be integral multiples of 100.*)
8. k=Number of nearest neighbors
9. numattr=Number of attributes
10. Sample[][]: array for original minority class samples

11. newindex: keeps a count of number of synthetic samples generated, initialized to 0
12. Synthetic[][]: array for synthetic samples (*Compute k nearest neighbors for each minority class sample only.*)
13. for i ← 1 to T
14. Compute k nearest neighbors for I, and save the indices in the nnarray
15. Populate(N, i, nnarray)
16. end for Populate (N,i, nnarray) (*Function to generate the synthetic samples*)
17. while N ≠ 0
18. Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.
19. for attr ← 1 to numattr
20. Compute: dif=Sample[nnarray[nn]] [attr] – Sample[i] [attr]
21. Compute: gap=random number between 0 and 1
22. Synthetic[new index][attr]+gap * dif
23. end for
24. newindex++
25. N=N – 1
26. end while
27. Return (*End of Populate*)

End of Pseudo-Code.

Variance captured while increasing feature space

In an effort to reduce the dimensionality further we opted to use principal components analysis (PCA) to choose enough eigenvectors to account for 0.95 of the variance of the sub-selected attributes [7]. We decided to standardize the data rather than center the data, which allows PCA to be computed by the correlation matrix rather than the covariance matrix. The maximum number of attributes to include through this transformation was 10. We then choose 0.95 for the value of variance covered. This allowed us to retain enough principal components to account for the appropriate proportion of variance. At the completion of this process we retained 288 components.

Automated feature selection methods

We utilized a supervised method to select features, a correlation-based feature subset selection evaluator [7]. This method of evaluation takes into account the value of a subset of features by analyzing the individual predictive ability of each feature along with the degree of sameness between them. The preference is to have low inter-correlation while having subsets of features that are highly correlated. Furthermore, we required that the algorithm iteratively add the highest correlated features with the class given there was not an existing feature in a subset that had a higher correlation with the feature being analyzed. We determined that we would search the space of features subsets using greedy hill climbing improved with a way of retracing. This retracing was governed by an environment of consecutive non-improving nodes. We set the direction of the search by starting with the empty set of attributes and searching forward. Additionally we specified that five

would be the number of consecutive non-improving nodes to allow before terminating the search. This method selected 19 attributes from the 60 features. Of those 19 features, only PT and GOS are original data inputs, representing approximately 11%; the other 89% are feature engineered (Table 3).

Evaluation and classifier

We used the receiver operator characteristic curves (ROC) which compare the false positive rate to the true positive rate. We can access the trade-off of the number of observations that are incorrectly classified as positives against the number of observations that are correctly classified as positives.

Area Under the Curve' (AUC) is the accuracy or total number of predictions that were correct,

$$\text{Accuracy} = \frac{\text{True positive} + \text{True Negative}}{\text{True Positive} + \text{False Negative} + \text{False Positive} + \text{True Negative}}$$

The misclassification rate or the error rate is defined as: **Error rate = 1 - accuracy**

We use other metrics in conjunction with the error rate to help guide the evaluation process,

namely Recall, Precision, False Positive Rate, True Positive Rate, False Negative Rate, and F-Measure [8].

Recall is the Sensitivity or True Positive Rate and demonstrates the ratio of cases that are positive and correctly identified,

$$\text{Recall} = \frac{\text{True positive}}{\text{True Positive} + \text{False Negative}}$$

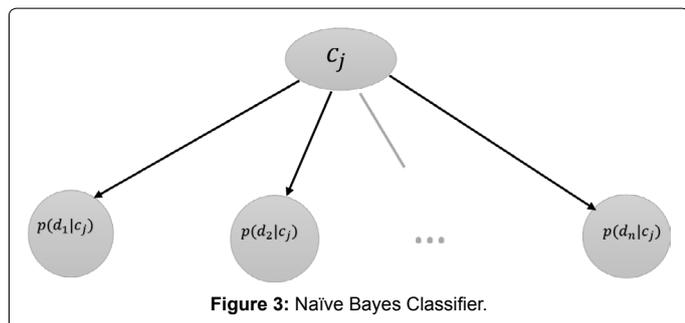
The False Positive Rate is defined as the ratio of cases that were negative and incorrectly classified as positive,

$$\text{False Positive Rate} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}}$$

The True Negative Rate or Specificity is defined as the ratio of cases that were negative and classified correctly,

Number of Folds(%)	Attribute
10	80th Percentile of A
10	Product of PI and PT
10	Sum of PR and GOS
10	PR Cubed
10	e pelvic tilt
10	e pelvic radius
10	e degree spondylolisthesis
30	PT
30	25 th Percentile of A
60	Quotient of PT and PI
70	Square root of PT
90	GOS
90	Range of elements in A
100	Standard Deviation of elements A
100	20 th Percentile of A
100	Sum of PR and GOS
100	Difference of PR and GOS
100	Quotient of PR and GOS
100	GOS Cubed

Table 3: Evaluation mode: 10 fold cross validation



$$F\text{-Measure} = 2(\text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}))$$

We simplified the task for classification by using a Naïve Bayes classifier which assumes attributes have independent distributions, and thereby estimate

$$P(d/c_j) = p(d_1 | c_j) \times p(d_2 | c_j) \times \dots \times p(d_n | c_j)$$

Essentially this is determining the probability of generating instance d given class c_j . The naïve bayes classifier is often represented as the following graph which states that each class causes certain features with a certain probability [9] (Figure 3).

In order to emphasize the benefits of the incorporation of feature engineering, feature selection, and PCA, we referenced prior research using two standard learning models and the rejsVM classifier [2]. All training and testing was uniformly applied as before.

Furthermore, we abandoned SVM as a base and instead choose to show the value of incorporating our methods within a simple Naïve Bayes algorithm [10-13]. Moreover, methods such as Feature Bayes may be used as a decision support tool for healthcare providers, particularly for those providers that have minimal resources or limited access to an ongoing professional peer network [14-16] (Tables 4 and 5).

Methods that produce high true positives and low false positives are ideal for medical settings. These allow healthcare providers to have a higher degree of confidence in the diagnoses provided to patients [17,18]. Given a small dataset, which is typical of biomedical datasets, feature Bayes helps to maximize the predictive accuracy that could benefit the medical expert in future patient evaluations [19,20] (Table 6).

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.855	0.115	0.883	0.855	0.869	0.935	Abnormal
	0.85	0.145	0.857	0.885	0.871	0.935	Normal
Weighted Avg.	0.87	0.13	0.87	0.87	0.87	0.935	

Table 4: Detailed accuracy by class (40% Train).

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.894	0.029	0.977	0.894	0.933	0.985	Abnormal
	0.971	0.106	0.872	0.971	0.919	0.985	Normal
Weighted Avg.	0.927	0.062	0.932	0.927	0.927	0.985	

Table 5: Detailed accuracy by class (80% Train)

Training Size	Method	Accuracy
40%	SVM (linear)	85
	SVM (KMOD)	83.9
	rejsVM (wr=0.04)	96.5
	Naïve Bayes (6-original data)	87.7
	Naïve Bayes (60-transformed data)	81.8
	Feature Bayes	93.5
80%	SVM (linear)	84.3
	SVM (KMOD)	85.9
	rejsVM (wr=0.04)	96.9
	Naïve Bayes (6-original data)	81.5
	Naïve Bayes (60-transformed data)	77.2
	Feature Bayes	98.5

Table 6: Comparison of the performance of different methods.

$$\text{True Negative Rate} = \frac{\text{True Negative}}{\text{False Positive} + \text{True Negative}}$$

The False Negative Rate is the proportion of positive cases that were incorrectly classified as negative,

$$\text{False Negative Rate} = \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}}$$

Precision is the ratio of the positive cases that were predicted and classified correctly,

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False Positive}}$$

F-Measure is computed using the harmonic mean and allows some average of the information retrieval precision and recall metrics. The higher the F-Measure value, the higher classification quality,

Conclusion

The analysis of the vertebral column data allowed us to incorporate feature engineering, feature selection, and model evaluation techniques. Given these new methods, we were able to provide a more accurate way of classifying pathologies. The feature Bayes method proved to be valuable by obtaining higher true positives and lower false positives than traditional or more current methods such as revo SVM. This makes it a useful method as a biomedical screening tool to aide healthcare providers with their medical decisions. Further studies should be developed surrounding the analysis of the feature Bayes method. Moreover, a comparison of ensemble learning techniques using feature Bayes could prove beneficial.

References

- Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD (2010) Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. The Journal of Machine Learning Research 11: 171-234.
- da Rocha Neto AR, Sousa R, Barreto GDA, Cardoso JS (2011) Diagnostic of pathology on the vertebral column with embedded reject option. Pattern Recognition and Image Analysis 6669: 588-595.
- Berthonnaud E, Dimnet J, Roussouly P, Labelle H (2005) Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters. Journal of spinal disorders & techniques 18: 40-47.
- Aghagolzadeh M, Soltanian-Zadeh H, Araabi B, Aghagolzadeh A (2007) A hierarchical clustering based on mutual information maximization. Image Processing 1: 1277-1280.
- Nguyen XV, Chan J, Romano S, Bailey J (2014) Effective global approaches for mutual information based feature selection. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic

-
- minority over-sampling technique. *Journal of artificial intelligence research* 16: 321-357.
7. Hall MA (1999) Correlation-based feature selection for machine learning. *CiteSeer_{SM}*.
 8. Powers DMW (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *CiteSeerx5M*.
 9. Zhang H (2004) The optimality of naive Bayes. *AA* 1: 3.
 10. Alba E, García-Nieto J, Jourdan L, Talbi EG (2007) Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. *Evolutionary Computation*.
 11. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, et al. (2015) Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports* 5: 10312.
 12. Brown G, Pocock A, Zhao MJ, Luján M (2012) Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research* 13: 27-66.
 13. Hand DJ, Yu K (2001) Idiot's Bayes—not so stupid after all? *International statistical review* 69: 385-398.
 14. Jordan A (2001) On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems* 14: 841-848.
 15. López FG, Torres MG, Batista BM, Pérez JAM, Moreno-Vega JM (2006) Solving feature subset selection problem by a parallel scatter search. *European Journal of Operational Research* 169: 477-489.
 16. Murty MN, Devi VS (2011) *Pattern recognition: An algorithmic approach*.
 17. Neto ARR, Barreto GA (2009) On the application of ensembles of classifiers to the diagnosis of pathologies of the vertebral column: A comparative analysis. *Latin America Transactions, IEEE (Revista IEEE America Latina)* 7: 487-496.
 18. Rennie JD, Shih L, Teevan, J, Karger DR (2003) Tackling the poor assumptions of naive bayes text classifiers. *ICML* 3: 616-623.
 19. Rish I (2001) An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, IBM, New York.
 20. Yamada M, Jitkrittum W, Sigal L, Xing EP, Sugiyama M (2014) High dimensional feature selection by feature-wise kernelized lasso. *Neural computation* 26: 185-207.