A Quotient Basis Kernel for the prediction of mortality in severe sepsis patients

Vicent Ribas Ripoll[1], Enrique Romero [1], Juan Carlos Ruiz-Rodríguez[2] and Alfredo Vellido [1] *

1- Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Edifici Omega, Campus Nord, 08034, Barcelona - Spain

2- Critical Care Department, Vall d'Hebron University Hospital, Vall d'Hebron Research Institute, UAB, Barcelona - Spain

Abstract. In this paper, we describe a novel kernel for multinomial distributions, namely the Quotient Basis Kernel (QBK), which is based on a suitable reparametrization of the input space through algebraic geometry and statistics. The QBK is used here for data transformation prior to classification in a medical problem concerning the prediction of mortality in patients suffering severe sepsis. This is a common clinical syndrome, often treated at the Intensive Care Unit (ICU) in a time-critical context. Mortality prediction results with Support Vector Machines using QBK compare favorably with those obtained using alternative kernels and standard clinical procedures.

1 Introduction

In this brief paper, we summarily describe a novel kernel for multinomial distributions, based on a suitable reparametrization of the input space through algebraic geometry and statistics. This kernel, named Quotient Basis Kernel (QBK), is the result of calculating the covariance of the design matrix of a Gröbner basis of the data. It has been shown that such a representation is very closely related to graphical models [1] in such a way that these kernels could be considered as *open-box* methods and thus comply with model interpretability requirements [2]. The downside is that the calculation of Gröbner bases is computationally costly.

The so-called QBK is applied here to data transformation prior to classification in a medical problem concerning the prediction of mortality in patients suffering severe sepsis. Sepsis is a common clinical syndrome, defined by the presence of both infection and Systemic Inflammatory Response Syndrome (SIRS). The acute stages of this condition are severe sepsis, which implies organ dysfunction, and the more severe septic shock and multiorganic failure [3, 4]. The mortality rates of sepsis are very high, ranging from 12.8% for sepsis to 45.7% for septic shock [5].

Severe sepsis is often treated at the ICU, which is a critical care environment. This is a time-critical context in which decision-making follows very specific protocols on the basis of well-defined quantitative indices. The definition of quantitative approaches to mortality prediction due to severe sepsis at the ICU is

*This research is partially funded by Spanish research project TIN2012-31377.

therefore a relevant research problem. This clinical environment further demands methods that are robust and straightforward to apply within its constraints.

Using soft-margin Support Vector Machines (SVM), the performance of the proposed QBK method in the prediction of mortality due to severe sepsis is compared to that obtained with a number of alternative generative kernels. It is also compared to a standard method currently used in clinical practice that is based on the basal APACHE II score [6] (i.e. through the automatic selection of a threshold). It is shown that the mortality prediction results using QBK compare favorably with those of the alternative methods.

2 Methods

2.1 Gröbner Basis

A Gröbner basis G is a subset of an ideal I in a polynomial ring R. For a system of polynomials P, G is an equivalence system that presents very useful properties. For example, any polynomial f is a combination of those in P iff the remainder of f with respect to G is 0 (i.e., the Gröbner basis is a sub-set generating the ideal $I(P)$). Here, the division algorithm requires an order of certain type on the monomials [7]. Moreover, the set of polynomials in a Gröbner basis has the same collection of roots as the original polynomials. This is particularly useful for the problem at hand since a set of points A can be considered as the set of solutions of the system of polynomials P. The definition of the Gröbner basis and some useful results for the definition of the QBK are given in the following.

Definition 1 *[7] Term Ordering: A term-ordering on R is an ordering relation $\succ_\tau$ (or τ or $\succ$) on the terms of R satisfying*

1. *$x^\alpha \succ 1$ $\forall x^\alpha$ with $\alpha \neq 0$ and*
2. *$\forall \alpha, \beta, \gamma \in \mathbb{Z}_+$ such that $x^\alpha \succ x^\beta$, then $x^\alpha x^\gamma \succ x^\beta x^\gamma$*

Definition 2 *[7] Let τ be a term ordering on R and f a polynomial in R. The leading term of f, $LT_\tau(f)$ is the largest term with respect to τ among the terms in f.*

Definition 3 *Ideal generated by a set of polynomials: The ideal generated by a set of polynomials F is the smallest ideal containing F. It is denoted $\langle F \rangle$.*

Definition 4 *[7] Gröbner Basis: Let τ be a term ordering on R. A subset $G = g_1, \ldots, g_t$ of an ideal I is a Gröbner basis of I with respect to τ iff*

$$\langle LT_\tau(g_1), \ldots, LT_\tau(g_t) \rangle = \langle LT_\tau(I) \rangle \tag{1}$$

where $LT_\tau(I) = \{LT_\tau(f) : f \in I\}$.

Theorem 1 *[7, 8] Given a term ordering, every ideal I except $\{0\}$ has a Gröbner basis and any Gröbner basis is a basis of I.*

Definition 5 *[7] Ideal of a set of support points: Let A be a set of unique support points $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$. The set $I(A)$ is the set of all polynomials whose zeros include the points in A.*

Definition 6 *Gröbner basis of unique points [7, 8]: Let A be a set of n unique points $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ and τ a term ordering. A Gröbner basis of A, $G = g_1, \ldots, g_t$, is a Gröbner basis of $I(A)$. Therefore, the points in A can be presented as the set of solutions of*

$$\begin{cases} g_1(\mathbf{a}) = 0 \\ g_2(\mathbf{a}) = 0 \\ \quad \cdots \\ g_t(\mathbf{a}) = 0 \end{cases} \tag{2}$$

2.2 Quotient Basis Kernel

Let us formally define the Quotient Basis EST_τ that shall be used in this short paper to define the QBK.

Definition 7 *[7] Quotient Basis:*

Let A be a set of unique support points $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ and τ a term ordering. A monomial basis of the set of polynomial functions over A is

$$EST_\tau = \{x^\alpha : x^\alpha \notin \langle LT(g) : g \in I(A)\rangle\} \tag{3}$$

As a consequence, EST_τ comprises the elements x^α that are not divisible by any of the leading terms $\mathrm{LT}(g)$ of the elements of the Gröbner basis of $I(A)$.

Theorem 2 *[7] The set EST_τ has as many elements as there are support points.*

Definition 8 *Design Matrix*

Let τ be a term ordering and let us consider an ordering over the support points $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$. We call design matrix (i.e. EST_τ evaluated in A) the following $n \times c$ matrix

$$Z = [EST_\tau]\big|_A \tag{4}$$

where c is the cardinality of EST_τ and n is the number of support points.

Theorem 3 *[7] Matrix Z is non-singular, and its covariance is a kernel.*

Definition 9 *Quotient Basis Kernel (QBK): The covariance of the design matrix of the quotient basis EST_τ of a a set of unique support points $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$,*

$$cov(Z) = E\left(Z - E(Z)\right)\left(Z - E(Z)\right)^t,$$

which is a kernel, is the QBK.

3 Results

3.1 Dataset description

This work resorts to a prospective observational cohort study of adult patients with severe sepsis. The study was conducted at the Critical Care Department of the Vall d'Hebron University Hospital (One of the main metropolitan hospitals in Barcelona, Spain), and was approved by the Research Ethics Committee of the Hospital. The resulting database includes data from 354 patients with severe sepsis, collected at the Vall d'Hebron ICU. Mortality for this population was 26.34%. In this study mortality was assessed through the following variables: numbers of dysfunctional organs (3.18 ± 1.32), presence of mechanical ventilation (66.71% of patients), severity measured through the APACHE II score (23.03 ± 9.62) and the Surviving Sepsis Campaign Resuscitation Bundles (i.e. adminstration of antibiotics with haemocultures in the first 6h of evolution, 31.41 % of the population [9, 10]).

3.2 Risk-of-Death Assessment with the APACHE II Mortality Score

The Risk-of-Death (ROD) formula based on the APACHE II score is a standard method in use in the critical care field. It can be expressed as [6]:

$$\ln\left(\frac{ROD}{1-ROD}\right) = -3.517 + 0.146 \cdot A + \epsilon, \tag{5}$$

where A is the APACHE II score and ϵ is a correction factor that depends on clinical traits at admission in the ICU. For instance, if the patient has undergone post-emergency surgery, ϵ is set to 0.613.

The application of this formula with a threshold of $\gamma = -0.25$ to the population under study yielded a classification error rate of 0.28, a specificity (true negative cases ratio, where a true negative is a patient correctly classified as *not being* at ROD) of 0.82 and a sensitivity (true positive cases ratio, where a true positive is a patient correctly classified as *being* at ROD) of 0.55. The corresponding area under the ROC curve (AUC) was 0.70.

3.3 Mortality Prediction with the Quotient Basis Kernel

The performance of the QBK was tested against other generative kernels, including the *exponential*, *centred* and *inverse* kernels described in [11]. For the sake of comparison, other well established kernels (*linear*, *polynomial* and *Gaussian*) were also tested. In order to improve the computation time for all kernels, the input data was first transformed into decile ranges. The QBK is calculated by taking the covariance after transforming the input data points with EST_τ. At this stage, it is important to note that the QBK accounts for all the interactions between the different input variables, which means that the 4 input variables are conditionally dependent [12] and that they can be represented with a fully connected graph. This interpretation is consistent with standard clinical practice.

The classification for all kernels was implemented using Matlab's Support Vector Machine QP solver from the Bioinformatics and Optimization Toolboxes. A grid search yielded that the most appropriate value for C parameter of the SVM was 10 for each kernel. A 10-fold cross-validation was used to obtain the classifiers, which were evaluated over a test dataset. The latter was obtained by random sampling 10% of the initial data before cross-validation.

The results summarized in Table 1 show that the QBK consistently produced the best outcome in terms of accuracy, specificity and AUC (which, in a way, summarizes sensitivity and specificity). The resulting sensitivity is comparable to the best results of alternative methods. It is also apparent that for this particular study, there are no major differences of performance between the other generative kernels (exponential, centred and inverse). The exponential and the centred kernels have the same accuracy as the Gaussian and polynomial. The latter, however, present slightly better sensitivities (i.e. 0.66 vs 0.65).

Despite the fact that the accuracy and specificity obtained with the APACHE II score are similar to those obtained with some of the kernels evaluated, the corresponding sensitivity is quite low. This poor sensitivity may be the result of the APACHE II score including non-sepsis specific clinical traits (for example, the performance of haemocultures, antibiotic administration or vasoactive drug administration).

4 Conclusion

The SVM classifiers in the reported experiments were trained with the transformed data resulting from the use of seven different kernels, out of which four were generative, while the rest were considered to be well-suited to the problem at hand. The investigated kernels provided accurate and medically actionable results, whilst keeping an acceptable balance between the different parameters of interest (accuracy rate, sensitivity and specificity).

The new kernel proposed in this paper, the QBK, is defined through the Gröbner basis of an algebraic ideal. It has been shown to outperform not only the alternative kernels, but also the clinical standard method based on the APACHE II score in the problem of mortality prediction for septic patients. In fact, all kernels outperform the standard APACHE II ROD formula in terms of accuracy, The QBK being the best according to all the criteria: accuracy, sensitivity and specificity.

Even though the QBK yields the best results, a word of caution must be given regarding its computation time. For high-dimensional datasets or very big input matrices (which is not the case of the analyzed data), the calculation of a Gröbner basis can be very time-consuming, even though the computational efficiency of the algorithms to calculate these bases has improved significantly over the last years.

Kernel	AUC	Error Rate	Sens.	Spec.
Quotient	0.89	0.18	0.70	0.86
Exponential	0.75	0.21	0.70	0.82
Inverse	0.62	0.22	0.70	0.82
Centred	0.75	0.21	0.70	0.82
Gaussian	0.83	0.24	0.65	0.81
Poly (order 2)	0.69	0.28	0.71	0.76
Linear	0.62	0.26	0.62	0.78
Apache II	0.70	0.28	0.55	0.82

Table 1: Results for SVM with diverse kernels and for the ROD formula based on the APACHE II score.

References

[1] M. Pachter, B. Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.

[2] A. Vellido, J.D. Martín-Guerrero, P.J.G. Lisboa, Making machine learning models interpretable. In proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2012), Bruges, Belgium, pages 163-172.

[3] American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference, Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis, *Critical Care Medicine*, 20:864-874, LWW, 1992.

[4] M.M. Levy, M.P. Fink, J.C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S.M. Opal, J.L. Vincent, G. Ramsay, 2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference, *Intensive Care Medicine*, 29(4):530-538, Springer, 2003.

[5] A. Esteban, F. Frutos-Vivar, N. Ferguson, O. Peñuelas, J.Á. Lorente, F. Gordo, T. Honrubia, A. Algora, A. Bustos, G. García, I. Rodríguez, Ruiz. R, Sepsis incidence and outcome: contrasting the intensive care unit with the hospital ward, *Critical Care Medicine*, 35(5):1284-1289, LWW, 2007.

[6] W.A. Knaus, E.A. Draper, D.P. Wagner, J.E. Zimmerman, APACHE II: A severity of disease classification system, *Critical Care Medicine* 13: 818-829, LWW, 1985.

[7] G. Pistone, E. Riccomagno, H.P. Wynn, *Algebraic Statistics: Computational Commutative Algebra in Statistics*, Chapman and Hall CRC, 2001.

[8] B. Giglio, E. Riccomagno, H. Wynn, Gröbner basis strategies in regression, *Journal of Applied Statistics* 27(7):923-938, Taylor & Francis, 2000.

[9] V.J. Ribas, A. Vellido, J.C. Ruiz-Rodríguez, J. Rello J., Severe sepsis mortality prediction with logistic regression over latent factors, *Expert Systems with Applications*, 39(2):1937-1943, Elsevier, 2012.

[10] V.J. Ribas, J.C. Ruiz-Rodríguez, A. Wojdel, J. Caballero-López, A. Ruiz-Sanmartín, J. Rello, A. Vellido, Severe sepsis mortality prediction with Relevance Vector Machines, In proceedings of the 33^{rd} *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC 2011), pages 100-103, IEEE, 2011.

[11] A. Agarwal, H. Daumé III, Generative kernels for exponential families, In G. Gordon, D. Dunson, M. Dudík, editors, proceedings of the 14^{th} *International Conference on Artificial Intelligence and Statistics* (AISTATS), pages 85-92, MIT Press, 2011.

[12] Drton M., Sullivant S., Algebraic statistical models, *Statistica Sinica*, 17:1273-1297, ISS, Academia Sinica, 2007.