

Decision Support in Attribute Selection with Machine Learning Approach

Wagner Arbex

Brazilian Agricultural Research Corporation — Embrapa
Juiz de Fora, MG, Brazil
wagner.arbex@embrapa.br

Fabyano Fonseca e Silva

Federal University of Viçosa — UFV
Viçosa, MG, Brazil

Marcos Vinícius Gualberto Barbosa da Silva

Brazilian Agricultural Research Corporation — Embrapa
Juiz de Fora, MG, Brazil

Fabrizio Condé de Oliveira

Federal University of Juiz de Fora — UFJF
Juiz de Fora, MG, Brazil

Luis Varona

University of Zaragoza — UNIZAR
Zaragoza, Spain

Rui da Silva Vermeque

Brazilian Agricultural Research Corporation — Embrapa
Juiz de Fora, MG, Brazil

Carlos Cristiano Hasenclever Borges

Federal University of Juiz de Fora — UFJF
Juiz de Fora, MG, Brasil

Abstract—This paper proposes a method to simultaneously select the most relevant single nucleotide polymorphisms (SNPs) markers — the attributes — for the characterization of any measurable phenotype described by a continuous variable using support vector regression (SVR) with Pearson VII Universal Kernel (PUK). The proposed study is multiattribute towards considering several markers simultaneously to explain the phenotype and is based jointly on a statistical tools, machine learning and computational intelligence.

Keywords—decision support; attribute selection; machine learning; SVR; computational modeling

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are an abundant form of genomic variation, which differ from rare variants [1] and the basic assumption for genome-wide association studies (GWAS) is that the evaluated characteristic can be explained from this type of marker.

The traditional approach is to evaluate which markers that have a high association with the phenotype through the *p*-value of beta linear regression between each SNP and the phenotype. After this step, the most relevant SNPs are analyzed for proximity to some region that is associated with that feature or other features that can be indirectly correlated with the phenotype in question.

Therefore, an alternative approach is to increase the number of markers, considering also those with small correlations on the trait. But, this fact creates two problems:

the number of markers is high and many of them are correlated. According to [2], such analysis requires the use of statistical methods that consider the selection of covariates – i. e., the multicollinearity problem -- and the regularization of the estimation process – i.e., the problem of dimensionality.

Other regression techniques were created to address this problem as ridge regression and partial least squares regression [3]. On the other hand, machine learning algorithms such as support vector machine (SVM) in GWAS considering multiple markers in classification problems, have demonstrated satisfactory performance as in [4], [5] and [6].

This study aims to propose a method that can simultaneously evaluate several SNPs in relation to the phenotype described by a continuous variable, unlike case-control dichotomous phenotypes addressed to the majority of GWAS studies. With this, there are two immediate benefits relative to standard methodology: one relating to the various levels of the phenotypes and the other by complex simultaneous interactions that may occur between the various markers.

To demonstrate the proposed method was used a sample of 343 samples (bulls) genotyped provided by the Brazilian Agricultural Research Corporation (Embrapa), and only 244 animals have female offspring, allowing the measurement of the phenotype evaluated.

SP 6753
P. 220

The genotype is the attribute set and was generated from the Illumina 56K chip, having a total of 56,947 SNP markers — the attributes.

In this study, the phenotype is the genetic potential of milk of an animal. This phenotype is computed from the milk production of their female offspring based on the methodology developed in [8]. The PTA milk is the “predicted transmitting ability”, being a measure of the expected performance of the daughters of the bull in relation to the average genetic herds [8].

Therefore, for instance, a 500 kg PTA for milk production means that if the bull is used in a population with genetic level same as to that used to evaluate it, each daughter will produce an average of 500 kg per lactation more than the average herd [8].

For the calculation of PTA milk, only the genetic effect is considered, eliminating all other environmental effects. Like this, the explanation of PTA from molecular marker information is consistent.

THE DATASET OVERVIEW

According to Table I, it was possible to notice a wide range of values of PTA, indicating the need of robust models for this mapping measurement through the genotype.

TABLE – STATISTICS OF MILK PTA (KG)

Minimum	1 ^o Quartile	Median	Mean	3 ^o Quartile	Maximum
-479.5	328.0	583.2	641.3	908.3	1,978.0

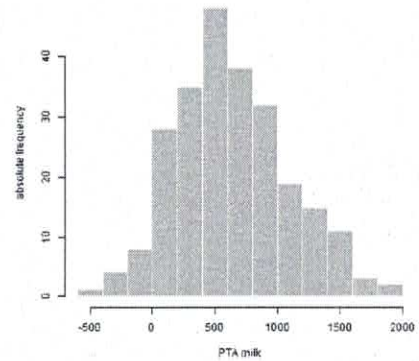
In the histogram of Figure 1, it was demonstrated that the distribution of the PTA for milk has positive skewness. From Figure 2, it is also noticed that there are two aberrant points higher than the average of the distribution of PTA.

FIRST SELECTION OF ATTRIBUTES

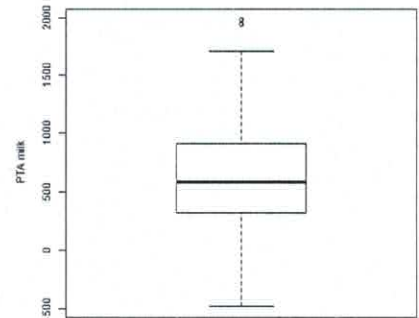
For comparison of the filters used to select the most important markers were created two databases: one without and one with quality control (QC).

There has been no standard preprocessing in the data without QC such as call-rate, minor allelic frequency (MAF) and Hardy-Weinberg equilibrium (HWE).

The purpose of this was not to eliminate SNPs with small effects alone, which when combined with other SNPs. They can be important in the description of PTA milk.



Histogram of PTA milk.



Boxplot of PTA milk.

In the sample without quality control, 6,192 markers showed no information due to errors in reading the chip Illumina Bovine 56k, and 3 markers had no allelic variation, so they were disregarded, totaling 50,752 for subsequent analysis.

For the construction of most significant groups of markers, we used the Spearman correlation coefficient. For the ranking of the effects of the markers, it was evaluated the Spearman correlation coefficient of each marker with the PTA and its corresponding *p-value*, as shown in Table II.

THE MACHINE LEARNING MODEL ON SVR

The support vector machine, SVM, is a machine learning technical for classify a set of data from dataset previously known and labeled. This label is described by a discrete variable.

The support vector regression or support vector machine with regression, SVR, is an extension of the SVM, where the discrete label is replaced with a variable that takes on continuous values.

TABLE – STATISTICS OF MILK PTA

Cluster	SNP selection	SNPs without QC	SNPs with QC
1	< 10-9	68	12
2	< 10-8	226	17
3	< 10-7	431	43
4	< 10-6	712	105
5	< 10-5	1.181	242
6	< 10-4	1.996	595
7	< 10-3	3.440	1397
8	< 10-2	6.512	3.356

The most significant difference between SVM and SVR approaches is the “construction of a tube” around the data to capture its variability and predictive power. This tube may take the linear and nonlinear forms depending on the kernel used.

The PUK, Pearson VII Universal Kernel or Pearson VII Function-based Universal Kernel, was adopted as the kernel for the method proposed. A Pearson VII function is able to easily change and adapt its two parameters σ and ω between the shapes of the Lorentzian and Gaussian function and even other functions [8].

Thus, this kernel has robustness as showing that percentage changes in parameters cause significant variations in the lower percentage in the RMSE of the predictions [8]. So, the PUK can replace the commonly applied kernels (linear, polynomial and RBF), possibly resulting in equal or superior performance regarding to the generalization of SVR [8].

The main advantage of this methodology is the choice of replacing the stock kernels by choosing the best parameters of PUK. This exchange of choice creates a clear gain, as each kernel is adopted in the training necessary to optimize its parameters and in the case of PUK, the kernel is not exchanged for another, because it mimics the behavior of other kernels.

For comparison of SVR models, was used 10-fold cross-validation on each of the 8 sets of data in Table II, and this procedure was repeated 10 times with different random seed for each partition created, for a total of 100 estimates for the Pearson correlation coefficient.

THE ATTRIBUTES SELECTION THROUGH SVR

Based on Cluster 8 (Table II) that generated the highest average and lowest standard deviation of the correlation, a wrapper, based on a binary genetic algorithm (GA) with fitness given by the cross-validation MSE is applied to a

second selection of markers. The entire methodology used in this second variable selection is based on [9].

As the number of combinations between 6,512 markers is extremely high, we used a GA to find the “best” markers in a skilled computational time, while not guaranteeing the uniqueness of the solution “optimal” found. Thus, the objective of the GA is to check the possibility of eliminating some of the best set markers generated from the first selection, as it is believed that the GA can better assess the interactions between markers than elimination made through filters quality control standards such as call-rate, HWE, MAF and LD.

The parameters adopted for the GA used for selecting SNPs were: probabilities of crossover and mutation equal to 0.6 and 0.033 respectively, population size and number of generations equal to 20.

DISCUSSION AN RESULTS

The first analysis was related to the accuracy of SVR models with linear, RBF and Pearson VII Universal kernels. From Table III, realize that the best model with PUK, both the lowest average and the lowest standard deviation of the correlation coefficient was set at *p-value* less than 10^{-2} .

TABLE – MEAN AND STANDARD DEVIATION OF THE CORRELATION COEFFICIENT OF PEARSON IN 10-FOLDS CROSS-VALIDATION WITH 10 REPETITIONS OF 3 SVR MODELS WITHOUT QC.

Cluster	SNP selection	SNPs	Linear	RBF	Pearson VII
1	< 10-9	68	0.60 (0.14)*	0.68 (0.11)	0.68 (0.11)
2	< 10-8	226	0.48 (0.17)	0.72 (0.09)	0.72 (0.09)
3	< 10-7	431	0.44 (0.16)	0.74 (0.09)	0.75 (0.08)
4	< 10-6	712	0.71 (0.09)	0.77 (0.08)	0.74 (0.09)
5	< 10-5	1.181	0.76 (0.09)	0.76 (0.08)	0.78 (0.08)
6	< 10-4	1.996	0.78 (0.08)	0.74 (0.08)	0.78 (0.08)
7	< 10-3	3.440	0.80 (0.08)	0.67 (0.13)	0.80 (0.08)
8	< 10-2	6.512	0.81 (0.08)	0.81 (0.08)	0.81 (0.08)

* Standard deviation of the estimates of the cross-validation.

Three models SVR based on the markers of the Cluster 8 of the base with QC showed equivalent prediction and accuracy, indicating a correlation averaging 0.80 with a standard deviation equal to 0.08 (Table IV).

Moreover, it seems that the Cluster 8 markers have a linear association with PTA milk because both the kernels RBF and PUK replicated this behavior.

The Table V shows that the subset of markers extracted from Cluster 8 showed higher mean correlation 0.84 with a standard deviation slightly lower 0.07 for the PUK.

This shows a significant gain in the use of GA for selecting the most informative SNPs without quality control. However, in the database with QA, there was a small increase in the mean correlation and was kept the same standard deviation 0.08.

TABLE – MEAN AND STANDARD DEVIATION OF THE CORRELATION COEFFICIENT OF PEARSON IN 10-FOLDS CROSS-VALIDATION WITH 10 REPETITIONS OF 3 SVR MODELS WITH QC.

Cluster	SNP selection	SNPs	Linear	RBF	Pearson VII
1	< 10-9	12	0.67 (0.10)*	0.67 (0.10)	0.67 (0.10)
2	< 10-8	17	0.64 (0.10)	0.67 (0.10)	0.67 (0.10)
3	< 10-7	43	0.59 (0.12)	0.68 (0.09)	0.70 (0.08)
4	< 10-6	105	0.31 (0.18)	0.72 (0.07)	0.71 (0.07)
5	< 10-5	242	0.67 (0.09)	0.77 (0.08)	0.78 (0.07)
6	< 10-4	595	0.77 (0.08)	0.69 (0.09)	0.79 (0.07)
7	< 10-3	1.397	0.78 (0.08)	0.79 (0.09)	0.79 (0.08)
8	< 10-2	3.357	0.80 (0.08)	0.75 (0.08)	0.80 (0.08)

* Standard deviation of the estimates of the cross-validation.

TABLE – MEAN AND STANDARD DEVIATION OF THE CORRELATION COEFFICIENT OF PEARSON IN 10-FOLDS CROSS-VALIDATION WITH THE BEST SUBSET FOUND ON CLUSTER 8.

Wrapper	Kernel		
	Linear	RBF	PUK
GA without QC	0.84 (0.07)	0.67 (0.14)	0.84 (0.07)
GA with QC	0.82 (0.08)	0.44 (0.25)	0.81 (0.08)

In relation to the work of [4] and [5], is possible to observe an improvement in using the PUK over other kernels analyzed on a them, as [5] uses the linear and RBF kernels with default values in R software, that is, the parameters SVM have not been optimized.

In the case of [4], only the RBF kernel was studied, and to find the “best” parameters C and γ , there was an extensive search grid.

However, neither of the two studies extracts from groups, constructed from the p -value, the “best” subset explanation for the phenotype and this was accomplished by applying the technique of variable selection wrapper based on the prediction error of the SVR as fitness GA.

The PUK proved to be robust to capture the behavior of linear and RBF kernels, as long as the appropriate parameters are used. Therefore, on the method proposed in this paper, it is necessary to assess the performance of the SVR with PUK.

However, regardless the kernel adopted, the mathematical formulation of the SVR brings a disadvantage regarding the biological interpretation, it is not possible to directly assess the optimal hyperplane and supports vectors which are the isolated effects of each marker and what are the markers that are simultaneously and the overall impact of this set.

The method developed in this study indicated that the 68 most significant markers of Cluster 1 without QC (Table III), has low predictive power and low accuracy compared to PTA milk and the subgroup with 3,073 markers showed high prediction and greater accuracy.

This may indicate that the PTA milk is a phenotype that is influenced by several markers with small effects on it, besides the possibility of epistasis and dominance, however, such genetic effects cannot be proven by the method suggested in this work.

The SVR model with PUK of groups 8 with and without QC showed high predictive power even in the presence of non-normality in the dependent variable PTA for milk, and have similar performance, and accuracy.

However, when applied the filter of the GA, the best subset generated from Cluster 8 without QC was higher both in prediction and accuracy as compared to group 8 with QC. This fact seems to show that the Cluster 8 markers without QC has sufficient and necessary for the explanation of the phenotype and the Cluster 8 with QC has markers necessary, but not sufficient.

Another point to be analyzed in depth which is subsequently SNPs that were not eliminated in the base without QC and can bring a high level of noise through the imputation used, inflating the explanatory power of SVR in 8 groups.

CONCLUSIONS AND FUTURE WORKS

The method developed in this work demonstrated robustness, because the initial set of markers without QC was composed of approximately 50,752 markers and reached up to 3,073 at the end of the selection process, ensuring good accuracy and high accuracy for the SVR model with PUK.

In addition to this fact, the GA able to eliminate most of the redundancy in the free base and QC on a smaller scale base with QC.

However, the remaining issue is to understand what level of redundancy that should remain the linkage disequilibrium between markers and this can be exploited by analyzing other cutting edges.

The standard filters — call-rate, MAF and HWE — used in the base with QC seem to delete markers essential to

explaining the phenotype PTA milk. From this study it is necessary to understand which filters are responsible for eliminating the most relevant SNPs.

The results obtained are promising for the application in GWAS, because most of the works in this area apply standard filters for preprocessing the database.

Moreover, the method proposed here can be the core for genomic selection aiming at predicting the breeding value of the individual from the genotype.

Future work are required, for instance, to determine fundamental physical map in which distances between 3,073 markers to verify that many of the markers are found indicating the same region or distinct regions in the genome, but this will be accomplished.

Furthermore, to verify the efficiency of the method developed here is required application in other databases SNPs associated with different phenotypes.

ACKNOWLEDGMENT

The authors thanks to reviewers who gave useful comments, and would like to express thanks to the National Research Center of Dairy Cattle (Embrapa Dairy Cattle) of Brazilian Agricultural Research Corporation (Embrapa) for providing database and the provision of the necessary infrastructure to conduct this work; to the Graduate Program in Computational Modeling of Federal University of Juiz de Fora (UFJF) for the academic support; and to the Coordination for the Improvement of Higher Level Personnel (CAPES) and the State of Minas Gerais Research Support Agency (FAPEMIG) for the support for the accomplishment of this paper.

REFERENCES

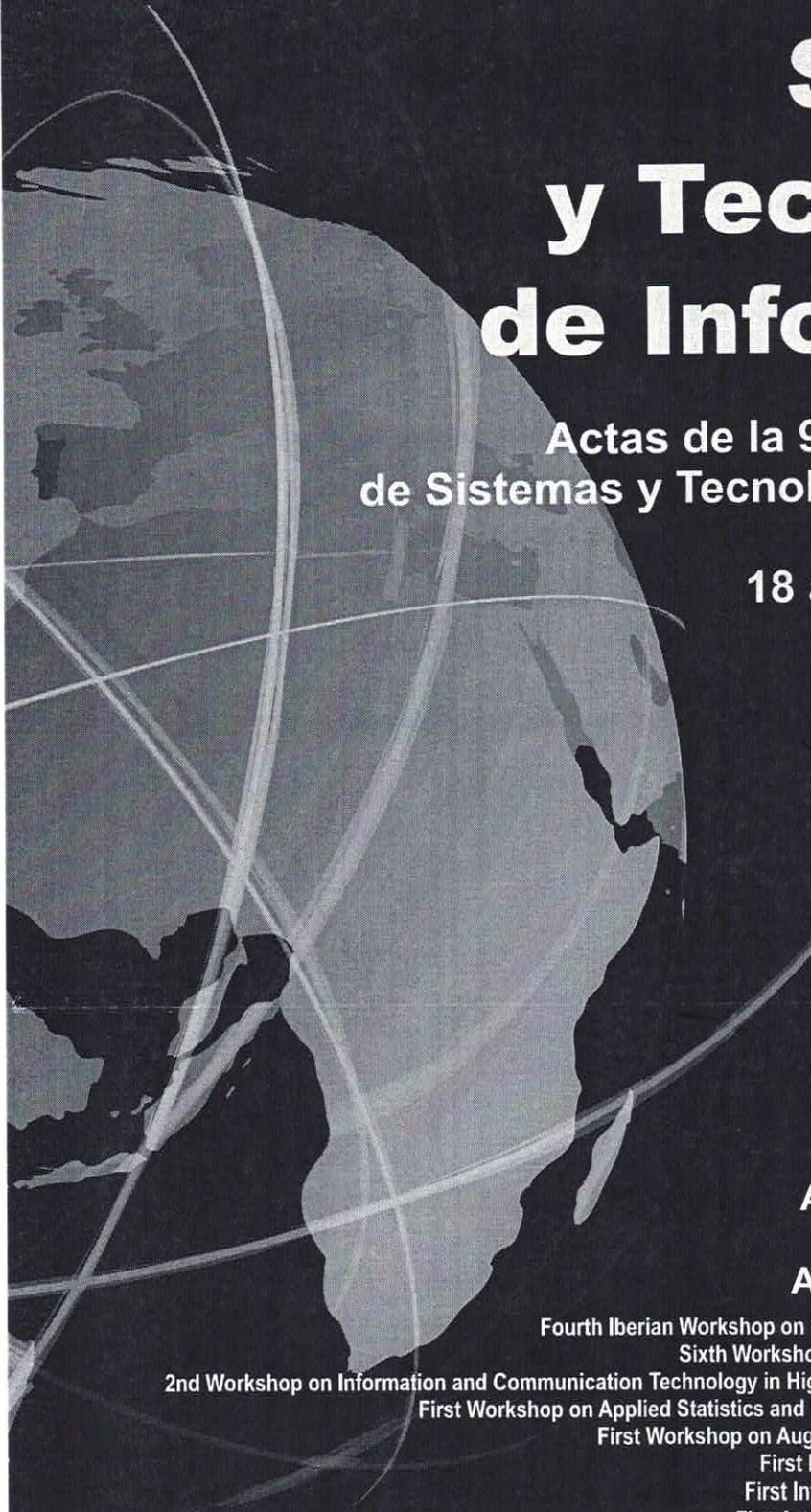
- A. J. Brookes. The essence of SNPs. *Gene*, vol.2, no. 234, pp. 177-186, July 1999.
- D. Gianola, M. Perez-Enciso, M. A. Toro (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics*, v.163, p. 347-365.
- G. Moser, M. S. Khatkar, B. J. Hayes, Herman W. Raadsma. (2010). Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. *Genetics Selection Evolution*. 42:37.
- F. Mittag, F. Büchel, M. Saad, A. Jahn, C. Schulte, Z. Bozhanovits, J. Simón-Sánchez, M. A. Nalls, M. Keller, D. G. Hernandez, J. R. Gibbs, S. Lesage, A. Brice, P. Heutink, M. Martinez, N. W. Wood, J. Hardy, A. B. Singleton, A. Zell, T. Gasser and M. Sharma (2012). Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. *Hum. Mutat.*, 33: 1708–1718. doi: 10.1002/humu.22161.
- Z. Wei, K. Wang, HQ. Qu, H. Zhang, J. Bradfield, C. Kim, E. Frackleton, C. Hou, JT. Glessner, R. Chiavacci, C. Stanley, D. Monos, SF. Grant, C. Polychronakos, H. Hakonarson (2009) From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. *PLoS Genet* 5(10): e1000678. doi:10.1371/journal.pgen.1000678.

HJ. Ban, J. Y. Heo, KS. Oh, KJ. Park. (2010). Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. *BMC Genetics* 11:26.

Babcock Institute. Genetic evaluation of dairy cattle in the USA. Available at <<http://babcock.wisc.edu/node/186>>.

B. Ünstü, W.J. Melssen, L.M.C. Buydens (2006). Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems* 81, 29 – 40.

D. E. Goldberg (1985). Genetic algorithms in search, optimization, and machine learning. 1.ed.. Publisher: Addison-Wesley Professional.



Sistemas y Tecnologías de Información

Actas de la 9ª Conferencia Ibérica
de Sistemas y Tecnologías de Información
Barcelona, España
18 al 21 de junio de 2014

Vol. I – Artículos

Editores
Álvaro Rocha
David Fonseca
Ernest Redondo
Luís Paulo Reis
Manuel Pérez Cota

Artículos de la Conferencia

Artículos de los Workshops

Fourth Iberian Workshop on Serious Games and Meaningful Play (SGamePlay)
Sixth Workshop on Intelligent Systems and Applications (WISA)
2nd Workshop on Information and Communication Technology in Higher Education: Learning Mathematics (TICAMES)
First Workshop on Applied Statistics and Data Analysis using Computer Science (ASDACS)
First Workshop on Augmented Reality and Wearable Computing (ARWC)
First International Workshop on Internet of Things (IoT)
First International Workshop on ICT for Auditing (WICTA)
First International Workshop on Learning Analytics (WLA)


aisti

Associação Ibérica de Sistemas
e Tecnologias de Informação

Salle

Universitat Ramon Llull

Credibilidad borrosa para mezclar diferentes fuentes de datos en la evaluación del riesgo operativo <i>Alejandro Peña, Isis Bonet, Christian Lochmuller, Alejandro Patiño</i>	240
CRUDi Framework Application - Insurance Company Case Study <i>Jorge Pereira, José Martins, Ramiro Gonçalves, Vítor Santos</i>	246
Database Synchronization Model for Mobile Devices <i>João Domingos, Nuno Simões, Paulo Pereira, Catarina Silva, Luís Marcelino</i>	252
Decision in Attribute Selection with Machine Learning Approach <i>Wagner Arbex, Fabrizzio Oliveira, Fabyano Silva, Luis Varona, Marcos Vinícius Silva, Rui Verneque, Carlos Cristiano Borges</i>	259
Decisive Factors for Adoption of Technology in e-Government Platforms <i>Marlon Freire, Nuno Fortes</i>	264
Denial of Service Attacks: An Overview <i>Vinko Zlomislíć, Krešimir Fertalj, Vlado Sruk</i>	270
Desarrollo de un método de conversión directa de una nube de puntos fotogramétrica, a puntos de movimiento de un cabezal de fresadora CNC <i>Galdric Santana Roma, Ayman Alitany</i>	276
Determinación de la capacidad de procesos de software siguiendo el modelo de evaluación de procesos NMX-I-15504 aplicado en el modelo de referencia NMX-I-059 bajo la herramienta AURAP <i>Angélica Astorga Vargas, Johanna Morales Bustamante, Brenda Flores Rios, Jorge Ibarra Esquer</i>	283
Disciplina, Orden y Forma. Livio Vacchini: Dinámica entre la Construcción y la Arquitectura en la Casa Vacchini de Costa Tenero <i>Laia Vives Arnella</i>	289
Diseño de un instrumento pedagógico para la enseñanza de la mejora de procesos software <i>Maria Clara Gómez Álvarez, Gloria Piedad Gasca-Hurtado, Jose Antonio Calvo-Manzano Villalón, Tomás San Feliu Guilabert</i>	295
Diseño y evaluación de un entorno web para la gestión del conocimiento de Mejora de Procesos Software <i>Claudia I. Martínez Alcalá, Jose A. Calvo-Manzano, Magdalena Arcilla-Cobián</i>	302
Doce Desafio - Aplicativo para Controle e Monitorização do Diabetes Tipo 1 em Dispositivos Móveis	310