Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Support vector machines for novel class detection in Bioinformatics

Eduardo J. Spinosa and André C.P.L.F. de Carvalho

Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, São Carlos, SP, Brasil
Corresponding author: E.J. Spinosa
E-mail: ejspin@icmc.usp.br/ejspin@yahoo.com

Genet. Mol. Res. 4 (3): 608-615 (2005)
Received May 20, 2005
Accepted July 8, 2005
Published September 30, 2005

ABSTRACT. Novelty detection techniques might be a promising way of dealing with high-dimensional classification problems in Bioinformatics. We present preliminary results of the use of a one-class support vector machine approach to detect novel classes in two Bioinformatics databases. The results are compatible with theory and inspire further investigation.

Key words: Novelty detection, Support vector machines, Bioinformatics, Gene expression analysis, Machine learning

INTRODUCTION

The ability to detect a new class or sub-class can be a useful feature for an artificial intelligent learning system. Slight modifications in data distribution might indicate, for instance, the appearance of a new class of data, or a pattern modification in a class that has already been modeled. This ability is known as novelty detection (ND) (Marsland, 2003) or one-class classification (Schölkopf et al., 1999, 2001), in reference to the fact that the training of the system is done based on examples of a single class representing the normal pattern.

Different approaches to the problem of ND have been proposed (Marsland, 2003), using various artificial intelligence techniques. Most of them have been applied to classification problems, where the system learns from labeled examples in a training phase, and later, in a test phase, assigns a class to each new unlabeled example. We believe that an ND approach to Bioinformatics can contribute to the solution of a variety of classification problems, by allowing the system to identify pattern changes and novel classes.

In the present study, a support vector machine (SVM) approach to ND is applied to Bioinformatics. In the next section, the one-class approach is presented. Initial results obtained with two gene expression datasets are analyzed in the "Experiments" section.

ONE-CLASS SUPPORT VECTOR MACHINE

Support vector machines

SVM is a machine learning technique based on Vapnik's Statistical Learning Theory (Vapnik, 1995), which provides it with a strong mathematical background. It has a great capacity for generalization, a very important feature in learning algorithms. As a consequence, it is less susceptible to overfitting than other techniques, and it achieves better results when dealing with new examples. SVMs are efficient and, since the function being optimized is convex, do not have problems dealing with situations where there are local optimums.

But above all these qualities, SVMs have proven to be robust in high dimensions, which make them especially interesting for applications in which the datasets consist of few examples and a high number of attributes. This is the case of many complex problems in Bioinformatics, including the classification problem based on gene expression data treated herein.

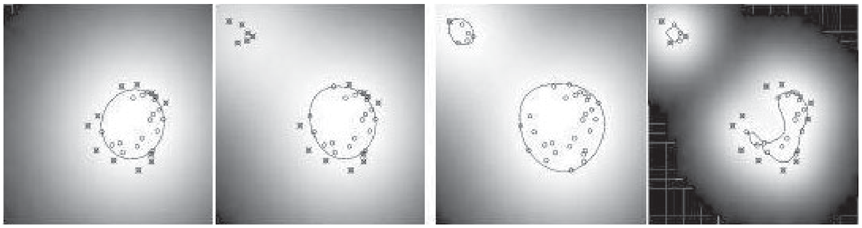
The one-class approach

The one-class approach was proposed by Bernhard Schölkopf (Schölkopf et al., 1999, 2001) and has been successfully applied to various problems (Campbell and Bennett, 2001; Manevitz and Yousef, 2001; Davy et al., 2002; Desobry and Davy, 2003).

Differentiation of members of novel and known classes is achieved by a data domain description. This is done by estimating a binary function that is positive where most of the data are located and negative elsewhere. A hyperplane with the largest possible margin is chosen to separate the training data from where the novel data are assumed to be.

The parameter ν (Nu) performs a trade-off between allowing more examples inside the description and making the description more general. This is an important setting that also makes the algorithm tolerant to noise that might be present in the training set. Figure 1 displays an example for different values of ν (Schölkopf et al., 1999).

Once the data domain description is known, the ND problem is reduced to a classification task where only one class exists (i.e., the normal class), thus the name one-class SVM. With this approach it is possible to detect examples in the test phase that are not well fitted to the data model that has been generated in the training phase.



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
frac. SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38
margin $\rho/\|w\|$	0.84	0.70	0.62	0.48

Figure 1. One-class support vector machine decision margins from Schölkopf et al. (1999), shown as an example.

EXPERIMENTS

We present initial results of the use of one-class SVM to detect novel classes in two Bioinformatics databases. The formulation that is used is implemented in a library called LIBSVM (Chang and Lin, 2004). Linear kernels were used in all experiments.

Databases

To test the performance of the one-class approach, the following databases were selected:

- Leukemia - identification of three types of leukemia (ALL-B, ALL-T and AML) from values of gene expression (Golub et al., 1999). The original database is composed of 72 examples (38 for training and 34 for testing) and 7,129 attributes (gene expression values).
- Lymphoma - distinction between germinal center and activated diffuse large B-cell lymphoma based on gene expression profiling (Alizadeh et al., 2000). The original database has 47 examples (34 for training and 13 for testing) and 4,026 attributes.

Methodology

Initially, all attributes were normalized to the interval [-1, +1]. Then, for each dataset, an experiment was carried out for each class considered a novelty.

In the Leukemia dataset for instance, there are three classes of the disease: ALL-B, ALL-T and AML. The original dataset is composed of two files, one for training and another for testing. Each file contains a few examples of each of the three classes. For the first experiment with this dataset, the goal is to identify the class ALL-B as novelty. In order to do that, the following three datasets were constructed from the two original ones:

- Training dataset (normal pattern), consisting of:
    - ALL-T examples from the original training dataset
    - AML examples from the original training dataset

- Testing dataset for the normal pattern, consisting of:
    - ALL-T examples from the original testing dataset
    - AML examples from the original testing dataset

- Testing dataset for the novel pattern, consisting of:
    - ALL-B examples from the original training dataset
    - ALL-B examples from the original testing dataset

Then, the same procedure was used for the remaining two classes (ALL-T and AML), generating two other experiments.

Therefore, for each class considered as novelty, three different datasets were created. The SVM was trained with the training set, which only contains examples of the normal pattern, and was then tested with both the normal testing set and the novel testing set.

According to the number of correct predictions, two accuracy rates were calculated: the normal accuracy rate measures how well the algorithm recognizes new examples of the known pattern, and the novelty accuracy rate does the same for examples of an unknown novel pattern.

A desirable situation is one in which the SVM algorithm is able to detect new patterns with high accuracy, but continues to classify normal examples with a good level of confidence as well. This tuning is made possible by the parameter ν, which determines how specific the data domain description should be, affecting the number of support vectors that are necessary. All experiments were performed for values of ν in the interval from 0.05 to 0.95, in 0.05 steps.

Analysis of the results

With these measures, it is possible to plot the curves of the accuracy rates of both the normal and the novel patterns with respect to the parameter ν. Figures 2, 3 and 4 display the results for the Leukemia database, each considering one of the three classes as novelty. The graphs include bars that show how many support vectors were used in each situation.

The first thing to notice is that the higher the value of the parameter ν, the higher the number of support vectors needed to describe the set of data. By increasing the value of ν, we are actually making the description more specific to the data in the training set (normal pattern), and this has a direct impact on the accuracy rates. The reason is that being more specific by increasing the parameter ν means gradually restricting the normal pattern to the examples of the training set and, consequently, making it harder for new normal instances to be recognized as such. This can be clearly seen in Figure 3, where the accuracy rate for the normal class decreases with a higher number of support vectors.

On the other hand, the more restricted the normal pattern becomes, the easier it is to

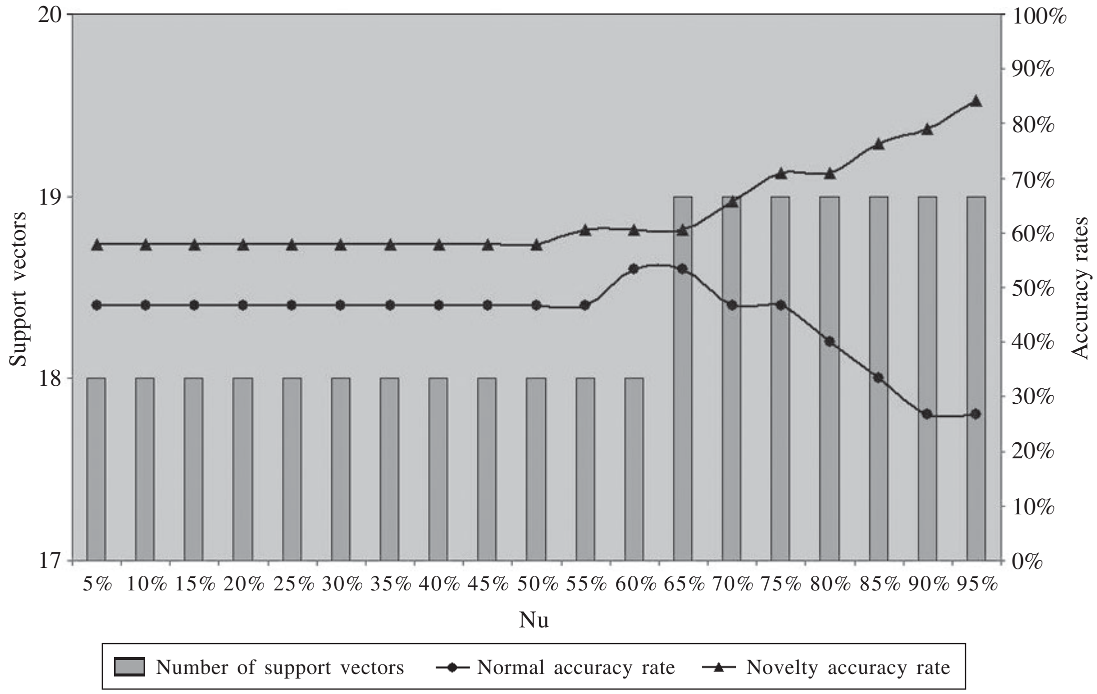


Figure 2. Accuracy rates for the Leukemia database considering ALL-B as novelty.

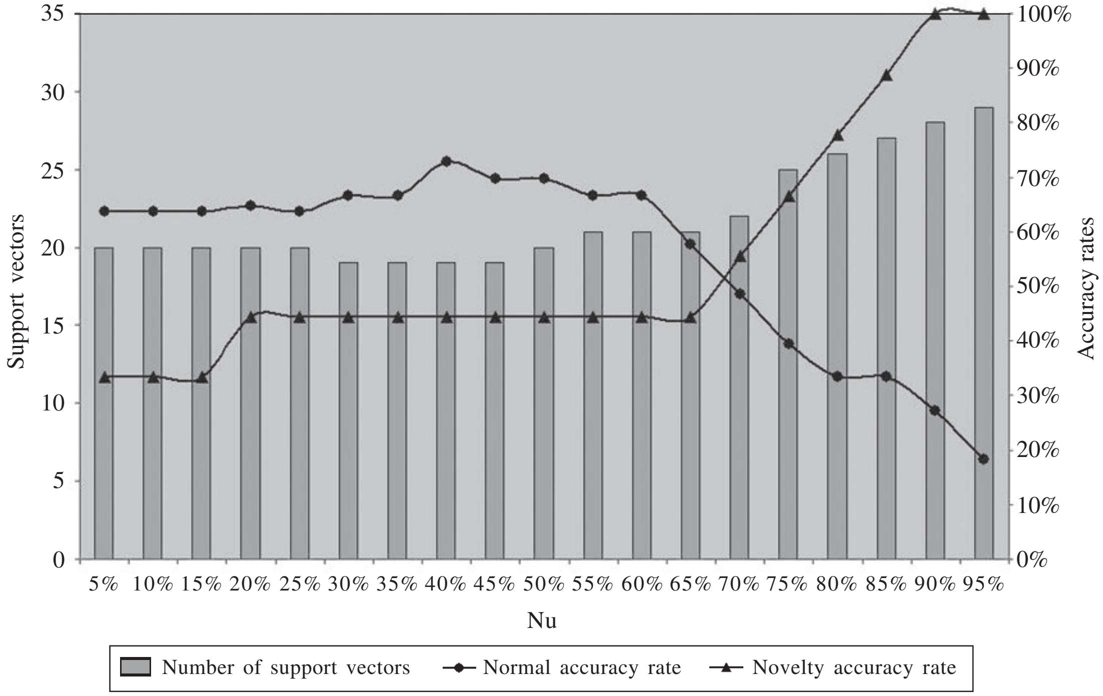


Figure 3. Accuracy rates for the Leukemia database considering ALL-T as novelty.

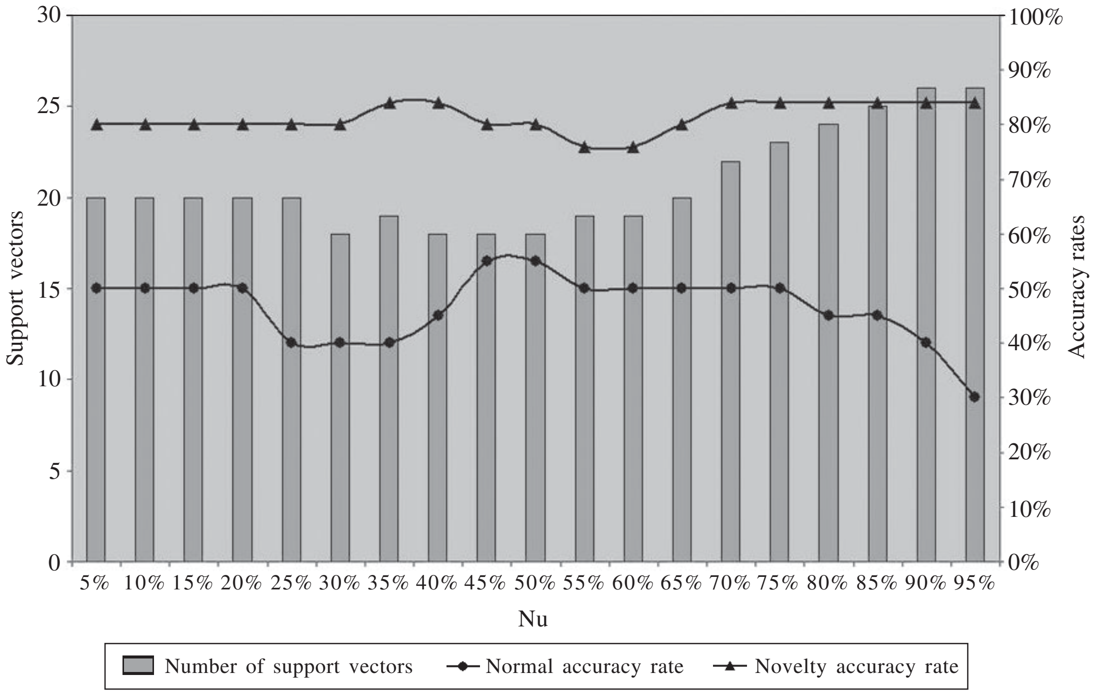


Figure 4. Accuracy rates for the Leukemia database considering AML as novelty.

have a new sample labeled as novelty. This contrary behavior can be seen in the constantly growing curves, which represent the accuracy rates for novelty.

We obtained similar results with the Lymphoma database (Figures 5 and 6).

In all experiments, the normal accuracy rate decreased and the novelty accuracy rate increased with growth, as parameter ν increased. In some situations (Figures 3, 5 and 6), these curves cross each other at a point where the value of ν could be considered optimal. But, even when there is no crossing (Figures 2 and 4), it is possible to identify the point where the best value for ν is achieved.

With regards to the levels of accuracy obtained in these experiments, and considering the high-dimensionality of both databases, the results are compatible. It is important to remember that the Leukemia database, for instance, is composed of 7,129 attributes and only a few examples are available for training and constructing the model (19 in the first experiment where ALL-B is considered novelty, 30 in the second where the novel class is ALL-T and 27 in the third experiment where the novelty is AML).

CONCLUSIONS

From the consistent results obtained in these early experiments, we believe that the use of SVMs to detect novel classes and perhaps even changes in the pattern of known classes has promise for Bioinformatics problems. Further experiments should be conducted with different learning algorithms and paradigms to allow performance comparisons with the SVM one-class approach. This is one of our future goals.

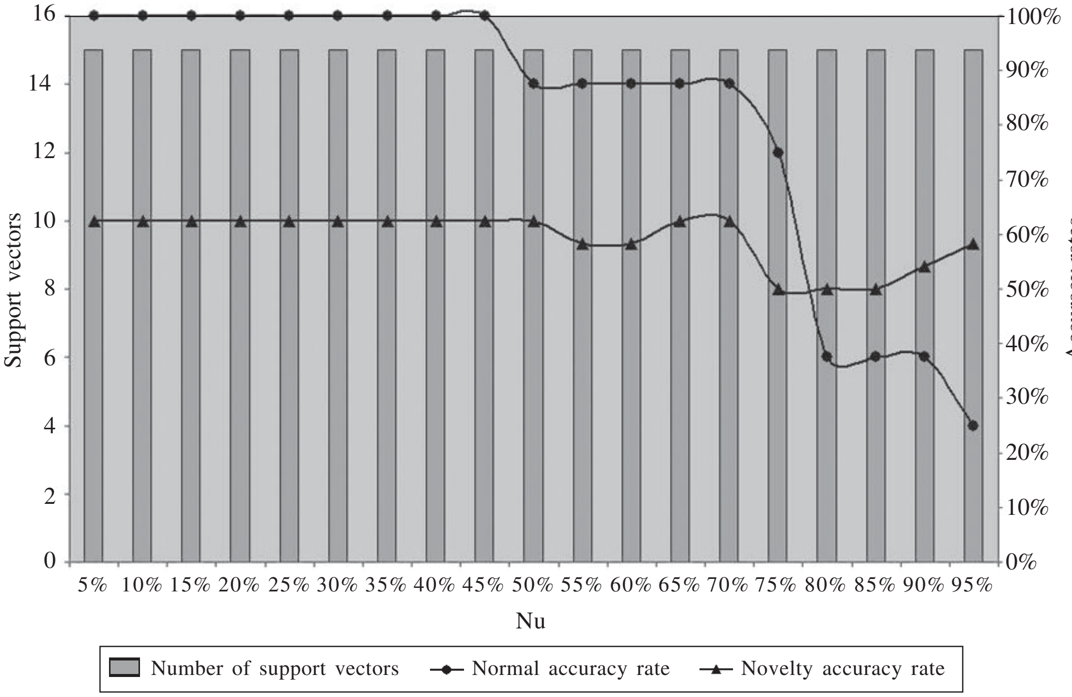


Figure 5. Accuracy rates for the Lymphoma database considering germinal center diffuse large B-cell lymphoma as novelty.

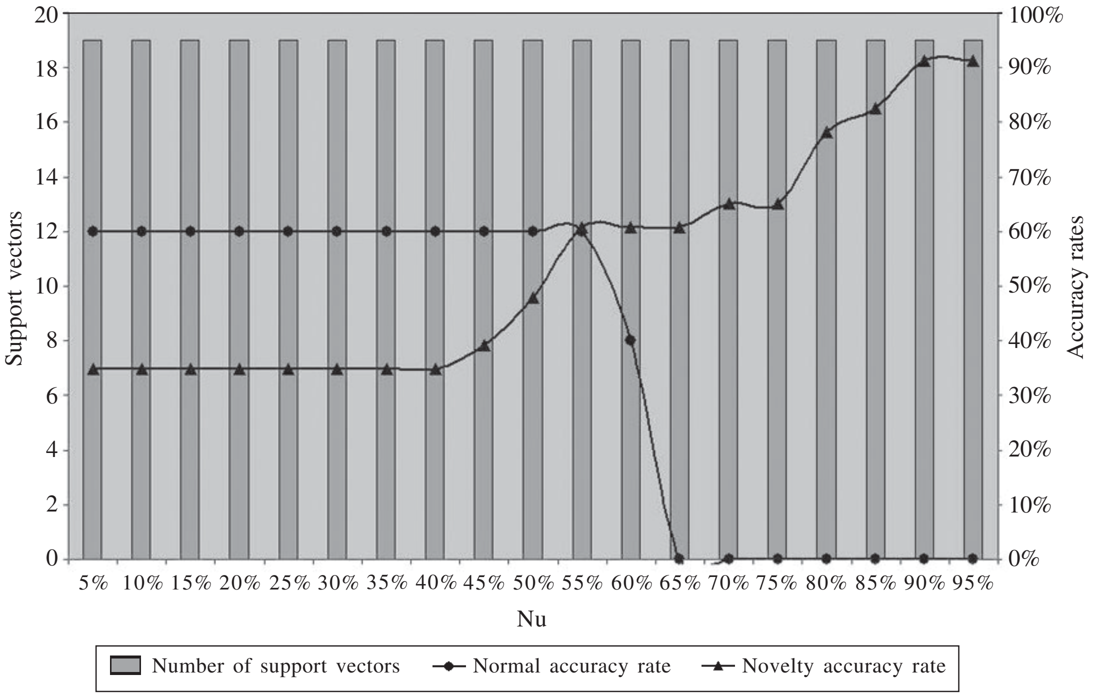


Figure 6. Accuracy rates for the Lymphoma database considering activated diffuse large B-cell lymphoma as novelty.

Bioinformatics presents a great number of new and interesting challenges to artificial intelligence researchers. Facing these challenges with robust tools and new approaches will lead to a higher level of understanding of multi-dimensional problems and to more effective ways to solve them.

ACKNOWLEDGMENTS

The authors thank CNPq for their financial support. CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) is a Brazilian government institution that promotes scientific and technological development.

REFERENCES

Alizadeh, A.A., Eisen, M.B., Davisintegral, R.E., Maintegral, C., Lossos, I.S., Rosenwaldintegral, A., Boldrick, J.C., Sabetintegral, H., Tranintegral, T., Yuintegral, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson Jr., J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and **Staudt, L.M.** (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature 403*: 503-511.

Campbell, C. and **Bennett, K.P.** (2001). A linear programming approach to novelty detection. *Adv. Neural Inf. Process. Syst. 14*: 395-401.

Chang, C. and **Lin, C.** (2004). *LIBSVM: a Library for Support Vector Machines*. National Taiwan University, Department of Computer Science and Information Engineering, Taiwan.

Davy, M., Gretton, A., Doucet, A. and **Rayner, P.J.W.** (2002). Optimised support vector machines for nonstationary signal classification. *IEEE Sig. Proc. Lett. 9*: 442-445.

Desobry, F. and **Davy, M.** (2003). Support vector-based online detection of abrupt changes. *Proc. ICASSP 4*: 872-875.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and **Lander, E.S.** (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression. *Science 286*: 531-537.

Manevitz, L.M. and **Yousef, M.** (2001). One-class SVMs for document classification. *J. Mach. Learn Res. 2*: 139-154.

Marsland, S. (2003). Novelty detection in learning systems. *Neural Comp. Surv. 3*: 157-195.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. and **Williamson, R.C.** (1999). Estimating the support of a high-dimensional distribution. Technical Report 87, Microsoft Research.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. and **Williamson, R.C.** (2001). Estimating the support of a high-dimensional distribution. *Neural Comp 13*: 1443-1471.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.