Statement of research interests

Vincenzo Russo (v.russo@pulsetech.it)

Department of Physics
Università degli Studi di Napoli "Federico II"
Complesso Universitario di Monte Sant'Angelo
Via Cinthia, I-80126 Naples, Italy

Overview

Machine learning is increasingly becoming a ubiquitous discipline, because there are a lot of application domains which need machine learning techniques nowadays: mathematical, physics and natural sciences as well as biological and medical disciplines are increasingly adopting machine learning techniques for data mining, patter recognition, forecasting, co-clustering etc. Business applications also needs machine learning instruments for various tasks, like market segmentation, market basket analysis, etc. Other disciplines which involve machine learning are image retrieval, computer vision, natural language processing, machine earing, anomaly detection, web mining, human-computer interaction, etc.

My research activity is just born by working on my master degree thesis [Rus08], which is currently the only research document issued.

In my master degree thesis I mainly worked on clustering techniques, providing some contributions to the *Support Vector Clustering* (SVC) [BHHSV01]. Such a work feed my research interests which are now focused in the **machine learning** world, with special attention on some application domains such as **information retrieval** and **knowledge management** as well as to my older interests in **Human-Computer Interaction (HCI)** and **web**.

I am currently looking for a position which allows me to combine some of the above interests.

Background

My research activity started with the work for the master degree thesis and was focused on clustering. I especially concentrated my own attention on SVC (and other support vector methods for clustering [Rus08, ch. 7]) providing some contributions. The SVC is a clustering technique which relies on Support Vector Machines (SVMs) and I also compared it with another cutting-edge clustering technique, the *Bregman Co-clustering* [BDG$^+$07].

The contributions were a better outliers handling, an enhanced kernel width

selection which can now work with all normalized kernels[1] and have a definitely lower computational complexity which makes it useful in practice. Moreover, an effective heuristics was added to the kernel selection, practically improving the selection of the kernel width values. Another effective heuristics was developed for the estimation of the second hyper-parameter of the SVC: the soft-margin parameter. Finally we provided another contribution for definitely making the SVC usable in practice: a new stopping criterion based on relative evaluation criteria (validity indices), which allows to exploit one of the key features of the SVC, i.e. the ability of auto-discovering the latent cluster structure.

The main application domains were **astrophysics data mining** (especially for testing missing data values robustness) and **text document clustering** (especially for testing the applicability of the SVC to this domains as well as its ability to handle sparse and very high-dimensional data). The SVC resulted robust w.r.t. missing data values, sparse data, very high dimensional data,[2] nonlinear separable problems, and arbitrary-shaped clusters. Moreover it has outliers handling ability, is application domain independent and is able to discover the cluster structure.

Current research

VONeural project. I am part of the VONeural project[3] team at the Department of Physics of University of Naples "Federico II". I have two roles: data mining algorithm researcher and designer and software engineer.

The VONeural project aims at developing a comprehensive data mining framework initially intended for astrophysics purposes, but including several instruments of general purpose. The enhanced version of the SVC (which I developed in my master thesis) will be also included in such a framework as well as the Bregman Co-clustering. The framework will be able to run on classical computers as well as on regular grid platforms and AstroGrid[4] platform.

Semi-supervised clustering. I am currently working on a semi-supervised clustering solution which combines SVC and classical SVMs; the experiments will run on text mining and astrophysics application domains. The idea follows by the SVC features: a first step performs the clustering process identifying the cluster structure of a subset including the most meaningful points.[5] A second step uses the identified cluster structure to train an SVM classifier.

[1] Originally the Secant-like kernel width generator was intended for Gaussian kernels only [LD05].

[2] The robustness w.r.t. missing data values, sparse data, and very high dimensional data as well as the applicability to the document clustering were shown for the first time in my master thesis [Rus08].

[3] See `http://people.na.infn.it/~astroneural/`.

[4] See `http://www2.astrogrid.org/`.

[5] The stable equilibrium points and/or support vectors.

Future research

Bregman Ball Vector Machines. I am preparing to develop an idea about future works on both SVC and Support Vector Machines. The idea consists of exploiting the Bregman divergences for developing a **new model of *vector machines***, which we could call ***Bregman Ball Vector Machines (BBVM)***. *Bregman divergences* [Bre67] form a large class of well-behaved loss functions which includes a number of well-known distances, such as the *Squared Euclidean Distance*, the *Mahalanobis* distance, the *KL-divergence*, the *I-divergence*,[6] etc. On the top of the Bregman divergences it is also possible to formulate the *Minimum Bregman Information Principle*, which generalizes famous principles such as the *Least Squares* principle and the *Maximum Entropy* principle.

The idea of developing new vector machines with Bregman divergences came out thanks to the *Minimum Enclosing Bregman Ball (MEBB)* problem [NN05, NN06], which is a generalization of the *Minimum Enclosing Ball (MEB)* problem. The latter is the problem underlying the SVC. In [NN05] we also found the generalization of the Bâdoiu-Clarkson (BC) algorithm, that is an approximation algorithm for efficiently solving the MEB problem. Such an extension is called Bregman-Bâdoiu-Clarkson (BBC) algorithm and is able to solve the MEBB problem in the same way the BC solves the MEB one.

BC algorithm is also used by Core Vector Machines (CVMs) [TKC05] and Ball Vector Machines (BVMs) [TKK07], reformulations of the classical SVMs as MEB problem. Since they rely on the BC algorithm and the latter was generalized to the BBC, it is likely to develop an extension of the CVMs/BVMs, the Bregman Ball Vector Machines (BBVMs), which works with the BBC and exploit the Bregman divergences.

Additionally, Bregman divergences could also be used for developing new type of kernels for classical SVMs and, probably, for extending another reformulation of the SVM, namely the Least Squares Support Vector Machines (LS-SVMs) [SGBBDM02]. Anyway, this has to be verified.

To conclude, here we have hypothesized about creating a new kind of vector machine, code-named Bregman Ball Vector Machine (BBVM). Since the MEB problem is the basic step of the Support Vector Clustering (and of other support vector methods for clustering [Rus08, chap. 7]), BBVM could have interesting implications for clustering applications too.

However, the feasibility and effectiveness of such an approach are far to be verified, but the results of the BVMs and the applicability of the MEBB problem are a first good point; it may worth to investigate a way of how combining them.

Information Retrieval, Knowledge Management, Web, HCI. Currently I have no way to go into these research fields thoroughly, but I hope to find such an opportunity someday. In the meantime I will study by myself.

[6] KL and I divergences promise good results handling text documents.

References

[BDG+07] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. A generalized Maximum Entropy approach to Bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8:1919–1986, August 2007.

[BHHSV01] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137, 2001.

[Bre67] Lev M. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[LD05] Sei-Hyung Lee and Karen M. Daniels. Gaussian kernel width generator for support vector clustering. In Matthew He, Giri Narasimhan, and Sergei Petoukhov, editors, *Advances in Bioinformatics and Its Applications*, volume 8, pages 151–162, 2005.

[NN05] Richard Nock and Frank Nielsen. Fitting the smallest enclosing Bregman balls. In *16th European Conference on Machine Learning*, number 3720 in Lectures Notes on Computer Science Series, pages 649–656. Springer-Verlag, 2005.

[NN06] Frank Nielsen and Richard Nock. On approximating the smallest enclosing Bregman Balls. In *SCG '06: Proceedings of the twenty-second annual symposium on Computational geometry*, pages 485–486, New York, NY, USA, 2006. ACM Press.

[Rus08] Vincenzo Russo. State-of-the-art clustering techniques: Support Vector Methods and Minimum Bregman Information Principle. Master's thesis, University of Naples "Federico II", Corso Umberto I, 80100 Naples, Italy, 2008. (Download from `http://thesis.neminis.org/2008/01/28/thesis-final-draft/`).

[SGBBDM02] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, and J. Vandewalle B. De Moor. *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore, 2002.

[TKC05] Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

[TKK07] Ivor W. Tsang, Andras Kocsor, and James T. Kwok. Simpler core vector machines with enclosing balls. In *Twenty-Fourth International Conference on Machine Learning (ICML)*, Corvallis, Oregon, USA, June 2007.