Reconocimiento Estadístico de Patrones
Máquinas de Soporte Vectorial y Series Temporales

Javier Giacomantone, Tatiana Tarutina, Verónica Artola, Armando De Giusti

Instituto de Investigación en Informática LIDI (III-LIDI)
Facultad de Informática – UNLP
{ jog, vartola,degiusti}@lidi.info.unlp.edu.ar, ttarutina@gmail.com

CONTEXTO

Esta línea de Investigación forma parte del Subproyecto "Tratamiento de imágenes digitales y video. Visión 3D", dentro del Proyecto acreditado "Algoritmos Distribuidos y Paralelos. Aplicación a Sistemas Inteligentes y Tratamiento Masivo de Datos" del Instituto de Investigación en Informática LIDI. Asimismo se integra con proyectos de cooperación bilateral con universidades del exterior y con un proyecto financiado por la Agencia Española de Cooperación Internacional (AECID). El planteo que se presenta se ha definido en la segunda mitad de 2008 y constituye una línea incipiente en el III-LIDI.

RESUMEN

Este trabajo describe una línea de I/D y los resultados esperados de la misma. El objetivo principal es estudiar, desarrollar y evaluar sistemas de reconocimiento automático de patrones en modo supervisado y no supervisado. En modo supervisado el objetivo principal es optimizar la generalización del clasificador. En particular son estudiados problemas caracterizados por medio de series temporales y clasificadores basados en máquinas de soporte vectorial (SVM). Los principales temas abordados son la selección de características, las técnicas de agrupamiento, el análisis de métricas y los métodos de optimización en SVM.

Palabras Clave: Reconocimiento de Patrones. Clasificación no lineal. Máquinas de Soporte Vectorial. Selección y extracción de características. Series temporales.

1. INTRODUCCION

Reconocimiento de patrones y en particular reconocimiento estadístico de patrones es un área de investigación interdisciplinaria tanto en la investigación básica de métodos fundamentales [1][2][3], como en sus aplicaciones [4][5]. El objetivo principal de un sistema de reconocimiento automático de patrones es descubrir la naturaleza subyacente de un fenómeno u objeto, describiendo y seleccionado las características fundamentales que permitan clasificarlos en una categoría determinada. Sistemas automáticos de reconocimiento de patrones permiten abordar problemas en informática, en ingeniería y en otras disciplinas científicas, por lo tanto el diseño de cada etapa requiere de criterios de análisis conjuntos para validar los resultados [6][7]. Las principales áreas de aplicación son, reconocimiento remoto, reconocimiento óptico de caracteres y escritura manuscrita, identificación de patrones en imágenes médicas, sistemas de clasificación en bioinformática, sistemas de identificación biométrica y clasificación de series temporales. Un modelo general de un sistema automático esta constituido por tres etapas, sensor, selector de características y clasificador. La primera etapa puede ser considerada a su vez como la que trata de obtener la representación más fiel del fenómeno estudiado, y un módulo que permite extraer las características del mismo. La línea de investigación propuesta está enfocada en la segunda y tercera etapa. Los métodos utilizados en reconocimiento de patrones se dividen en dos grandes categorías clasificación supervisada y clasificación no supervisada. El tipo de objetos o fenómenos considerados en esta línea de trabajo pueden ser descriptos por un conjunto de características numéri-

cas que definen patrones en un espacio n-dimensional. Por lo tanto el análisis de las distribuciones estadísticas de cada clase y los métodos de estimación de parámetros, permiten definir estrategias de diseño, evaluar y especificar los métodos de clasificación. La línea de investigación propuesta en este trabajo esta enfocada en el diseño de clasificadores basados en SVM y en la aplicación a problemas caracterizados por series temporales y la clasificación contextual de las mismas.

1.1 Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial (SVM) [8][9] son herramientas fundamentales en sistemas de aprendizaje automático, permitiendo el tratamiento de problemas actuales en reconocimiento de patrones y minería de datos tales como, reconocimiento y caracterización de texto manuscrito, detección ultrasónica de fallas en materiales, clasificación de imágenes médicas [10], sistemas biométricos [11], clasificación en bioinformática [12][13] y en física de altas energías [14]. Las SVM implementan reglas de decisión complejas, por medio de una función no lineal que permite mapear los puntos de entrenamiento a un espacio de mayor dimensión. En el nuevo espacio de características las clases son separadas por un hiperplano, siendo este el que maximiza la distancia entre el mismo y los puntos de entrenamiento. La distancia se denomina margen y esos vectores son los vectores de soporte. Las SVM cumplen un rol muy importante en teoría de aprendizaje estadístico y cuando es necesario entrenar un clasificador no lineal en un espacio de características de considerable dimensión con un número limitado de muestras. Podemos diferenciar dos aspectos importantes que en general reciben la denominación de maquinas de soporte vectorial, el uso de SVM en clasificación SVC y el uso de las mismas en regresión SVR [15]. La línea de investigación propuesta estudia ambos aspectos y en particular en el caso de clasificación mediante SVM tiene como objetivo diseñar sistemas con alta capacidad de generalización. Entre las tendencias actuales podemos mencionar las investigaciones sobre SVM paralelas y secuenciales (PSVM, SSVM) [16].

1.2 Series Temporales

Una serie temporal es una secuencia de puntos medidos a intervalos sucesivos, normalmente de tiempo, y en general a intervalos regulares. Las series temporales son el resultado de medidas de distintos fenómenos físicos en la naturaleza pero también son comunes en econometría, marketing, control industrial y como resultado de métodos de monitoreo y diagnóstico en medicina. Fundamentalmente su caracterización se da en el dominio espacial o en el dominio transformado de Fourier [17], Wavelets[18], Chirplet [19] y sus técnicas de análisis son fundamentalmente estadísticas y de procesamiento de señales. La clasificación de las series temporales obtenidas a partir de estudios funcionales del cerebro, son un ejemplo de abordaje multidisciplinario, donde uno de sus aspectos fundamentales es el de reconocimiento de patrones. La aplicación, adaptación y adecuada selección de kernels de SVM a series temporales es un tema de investigación actual [20][21]. La clasificación básica de las máquinas de soporte vectorial es binaria por lo tanto es importante el estudio de la extensión a multi-clasificación [22]. Las series pueden ser unidimensionales o multidimensionales con correlación tanto temporal como espacial. En el último caso se plantea un problema de reconocimiento de patrones complejo y de minería de datos (MDTSC multidimensional time series classification). Entre los temas actuales de investigación que involucran los conceptos anteriores podemos citar estudios en neurociencias por medio de resonancia magnética funcional (fMRI), electroencefalogramas (EEGs) y magneto-encefalogramas (MEGs) [23][24]. Los problemas anteriores requieren computo numérico intensivo demandando la especificación y desarrollo de sistemas paralelos para su implementación [25][26][27], debi-

do fundamentalmente a la complejidad y el volumen de datos procesados.

2. LINEAS DE INVESTIGACION y DESARROLLO

- Clasificación supervisada. Discriminadores lineales y no lineales.
- Métodos de estimación de parámetros para clasificadores Bayesianos.
- Clasificación no supervisada. Técnicas de agrupamiento (clustering).
- Selección y extracción de características.
- Métricas, pseudométricas y distancias ultramétricas en clasificación supervisada, no supervisada y selección de características.
- Criterios de evaluación de desempeño en sistemas de clasificación automática.
- Criterios y algoritmos para combinación de clasificadores.
- Maquinas de soporte vectorial. Kernels y algoritmos de optimización.
- Clasificación de series temporales y clasificación contextual.
- Caracterización y evaluación de la capacidad de generalización de los clasificadores propuestos.
- Paralelización y análisis de complejidad de los algoritmos propuestos

3. RESULTADOS OBTENIDOS /ESPERADOS

- ✓ Desarrollar modelos y optimizar algoritmos particulares de clasificación supervisada y no supervisada.
- ✓ Evaluación de los métodos de análisis de desempeño y su aplicación sobre los clasificadores y conjuntos de datos propuestos.
- ✓ Obtener mejoras y adecuar las técnicas de selección y extracción para el tratamiento de datos en espacios multidimensionales
- ✓ Dada la naturaleza interdisciplinaria de una línea de investigación como el reconocimiento de patrones, en particular en las áreas de aplicación, promover la integración entre las distintas líneas de investigación.
- ✓ Evaluar las técnicas propuestas sobre datos simulados y reales.
- ✓ Dada la naturaleza específica de las aplicaciones que implican cómputo intensivo para resolver las soluciones numéricas propuestas, transferir estos resultados para su investigación y posible tratamiento mediante técnicas de procesamiento paralelo y distribuido
- ✓ Transferir los resultados obtenidos, nuevas técnicas, algoritmos y tratamiento de datos experimentales de nivel fundamental a las áreas de aplicación principales.

4. FORMACION DE RECURSOS HUMANOS

En esta línea de I/D existe cooperación entre distintos subproyectos de investigación en el III-LIDI, fundamentalmente por la utilidad de los métodos estudiados para resolver problemas de clasificación en tratamiento masivo de datos, como una etapa fundamental en un sistema de visión por computador y por ser particularmente viables para su cómputo paralelo. En el marco de esta línea de investigación hay un investigador realizando su doctorado y se espera la realización de tesinas y tesis desarrollando aspectos particulares en sistemas automáticos de reconocimiento de patrones.

5. BIBLIOGRAFIA

1. Fukunaga K. “Introduction to Statistical Pattern Recognition”. Second Edition. Academic Press, 1990.
2. Devijer P. A., Kittler, J. “Pattern Recognition, A Statistical Approach”. Prentice Hall, 1982.
3. Batagelj V, Bock H, Ferligoj A. “Data Science and Classification”. Springer, 2006.
4. Devijer P, Kittler, J. “Pattern Recognition: theory and applications”. Springer, 1986.

5. Anke Meyer-Base. “Pattern Recognition for Medical Imaging”. Academic Press, 2004.
6. Kim H.Y., Giacomantone J. O., Cho, Z. H. Robust Anisotropic Diffusion to Produce Enhanced Statistical Parametric Map, Computer Vision and Image Understanding, v.99, p.435-452 (2005).
7. Kim H.Y., Giacomantone J. O., A New Technique to Obtain Clear Statistical Parametric Map by Applying Anisotropic Diffusion to fMRI, IEEE, International Conference on Image Processing. Proceedings, Genova, Italy,v.3, p.724-727 (2005).
8. Corte C, Vapnik V, Support vector networks. Machine Learning v.20, p.273-297 (1995).
9. Vapnik, V. The Nature of Statistical Learning Theory. N. Y. Springer (1995)
10. S. Li, T. Fevens, A. Krzyzak, S. Li. Automatic Clinical Image Segmentation Using Pathological Modelling, PCA and SVM, MLDN, LNAI 3587 pp.314-324, (2005).
11. Z. Lei, Y. Yang, Z. Wu. Ensembles of Support Vector Machine for Text-Independent Speaker Recognition, IJCSNS v.6 n.5A pp. 163-167, (2006).
12. Y. Li, J. Li. Predicting Subcellular Localization of Proteins Using Support Vector Machines with N-Terminal Amino Composition, ADMA 2005, LNAI 3584, pp. 618-625, (2005).
13. R. Boekhorst, I. Abnizova, L. Wernich. Discrimination of regulatory DNA by SVM on the basis of over- and under-represented motifs, ESANN pp. 481-486 (2008).
14. Vossen Anselm. Support Vector Machines in High Energy Physics, CERN, Geneva, Switzerland, pp.23-33 (2005).
15. Vapnik, V. Golowich S., Smola A. Support Vector Method for Function Approximation, Regression, Estimation and Signal Processing. In Advances in Neural Information Processing Systems, Vol 9, pag. 281-287. MIT Press, Cambridge, 1997.
16. L. Wang, M. Chang, J. Feng. Parallel and Sequential Support Vector Machines for Multi-label Classification, International Journal of Information Technology, v.11 n.9 pp. 11-18, (2005).
17. Grafakos Loukas. Classical and Modern Fourier Analysis, Prentice Hall. (2004).
18. D. B. Percival, A. T Walden. Wavelet Methods for Time Series Analysis, Cambridge University Press (2000).
19. J. R. Cui, etal. Time frequency análisis of visual evoked potentials using chirplet transform. IEE Electronic Letters v.41 p.n.4 pp.217-218 (2005).
20. S. Rüping. SVM kernels for time series analysis, G1-Workshop-Woche Lernen-Lehren-Wissen-Adaptivitet, pp.43-50 (2001).
21. K.Yang, C.Shahabi. A pca-based kernel for kernel pca on multivariate time series, IEEE Intern. Conf. on Data Mining (2005).
22. C. W. Hsu, C. J, Lin. A comparison of methods muti-class support vector machines, IEEE Trans. on Neural Networks v.13 pp. 415-425 (2002).
23. Javier Giacomantone, Armando De Giusti, ROC performance evaluation of RADSPM technique, XIV Congreso Argentino de Ciencias de la Computación (CACIC), Chilecito (2008).
24. W. A. Chaovalitwongse, P. M. Pardalos. On the Time Series Support Vector Machine using Dynamic Time Warping Kernel for Brain Activity Classification, Cybernetics and Systems Analysis v.44 pp.125-138 (2008).
25. N. Goddard, G. Hood, J. Cohen, W. Eddy, C. Genovese, D. Noll, L. Nystrom. Online analysis of functional mri datasets on parallel platforms. J.Supercomput., 11(3):295–318, (1997).
26. A. De Giusti, M. Naiouf. L. De Giusti, F. Chichizola. Dynamic Load Balancing on Non-Homogeneous Clusters. Lectures Notes in Computer Science, v.4330 pp.65-73. Springer Verlag, (2006)
27. T. Eickermann, W. Frings, F. Hossfeld, S. Posse, G. Goebbels. Supercomputer-enhanced functional mri of the human brain. IEEE Concurrency, 8(1):11–13, (2000).