Determinación de la Denominación de Origen de vinos chilenos basado en Máquinas de Soporte Vectorial

Anibal Rojas[1], Marco Mora[1] y Evelyn Villagra[2]

[1] Laboratorio de Investigaciones Tecnológicas en Reconocimiento de Patrones
Departamento de Computación, Universidad Católica del Maule, Chile
[2] Escuela de Ingeniería en Biotecnología, Universidad Católica del Maule, Chile
marcomoracofre@gmail.com
http://www.litrp.cl

Resumen Se presenta un método para determinar la denominación de origen de vinos chilenos basado en su concentración de metales. Se emplea un repositorio de 77 muestras de vinos y sus correspondientes concentraciones de metales. Se aplican dos funciones Kernel junto a clasificadores basados en Máquinas de Soporte Vectorial. Se comparan tres metaheurísticas para encontrar los hiperparámetros óptimos de los clasificadores. Para entrenarlos se aplica Validación Cruzada Dejando Uno Fuera. Los resultados se calculan en base al error promedio de las clasificaciones. Los porcentajes de error se estiman no superiores al 15 %, destacando la combinación de Recocido Simulado y Kernel Lineal como la más óptima.

Palabras Claves: Denominación de Origen, Concentración de Metales, Máquinas de Soporte Vectorial, Vinos Chilenos, Metaheurística

1 Introducción

En la investigación científica se ha aplicado la clasificación de datos para solucionar problemas, adquirir conocimientos y automatizar procesos, logrando resultados cada vez más precisos. Un campo de estudio que ha ganado auge es la industria del vino, dada su importancia como bien económico, clasificándose a través de su semilla [1], color [2, 3] y presencia de químicos [4]. Además, se establecen experimentos para determinar su denominación de origen tanto en Latinoamérica [5] como en el mundo [6], [7], [8]. En este contexto surgen las Máquinas de Soporte Vectorial (SVM) para clasificar datos complejos, uniéndose a otras técnicas para mejorar sus resultados ([9]), aunque requieren configurar sus datos de entrada e hiperparámetros. Para obtener valores óptimos se han revisado las técnicas aplicadas [10], mejorado las existentes [11] y aplicado ideas nuevas [12], [13], destacando a las metaheurísticas, que registran un espacio de búsqueda en un tiempo razonable y hallan buenas soluciones [14–20].

En este paper se desarrolla una comparativa de tres metaheurísticas para definir hiperparámetros óptimos aplicados en clasificadores basados en SVM y determinar la denominación de origen de vinos chilenos según sus índices de concentración de metales.

Responder a esta interrogante causaría un impacto en la industria del vino chileno, su elaboración, transacciones comerciales y aumentando su calidad y prestigio.

2 Materiales y Métodos

2.1 Descriptores para la denominación de origen

El repositorio de datos consiste en características para 77 muestras de vinos chilenos que incluye los niveles de concentración de Sodio, Potasio, Magnesio, Calcio, Hierro y Zinc.

2.2 Estructura del clasificador

Los experimentos emplean 77 SVM de salida binaria para cada denominación de origen. Para optimizar los hiperparámetros, éstas se ejecutan bajo Kernel Lineal y Función de Base Radial (RBF). Esto se repite para cada metaheurística, siendo ejecutado seis veces. Como espacio de búsqueda se tiene una grilla de valores entre $[-5, 5]$.

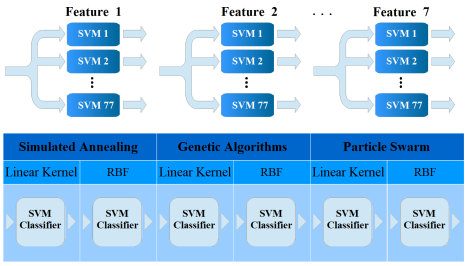

Fig. 1. Estructura del clasificador para cada característica y en forma general.

2.3 Entrenamiento de SVM y obtención de los resultados

Existen 539 conjuntos de entrenamiento para cada experimento. Las clases asignadas a los datos son (1) para la denominación de origen de un vino y (-1) para el resto. Se aplica Validación Cruzada Dejando Uno Fuera (LOOCV). Para cada iteración se genera un error que es el promedio de los errores particulares. El Error General se obtiene al detectar la iteración que ha generado el mínimo error promedio. Las salidas se calculan en base al Porcentaje de Error en la clasificación en vez del Porcentaje de Acierto.

2.4 Mediciones y experimentaciones realizadas

Se llevan a cabo seis experimentos: uso de Recocido Simulado con Kernel Lineal y RBF; uso de Algoritmos Genéticos con Kernel Lineal y RBF; uso de Optimización por Enjambre de Partículas con Kernel Lineal y RBF.

3 Resultados

Con Recocido Simulado y Kernel Lineal los porcentajes de error no superan el 10 %, llegando al 9.46 %; en RBF los niveles de error superan el 12 %. Con Algoritmos Genéticos y Kernel Lineal oscilan alrededor del 11 % de error y un experimento llega al 9 %; en RBF se observan niveles que llegan al 15 % de error promedio. Con Optimización por Enjambre de Partículas y Kernel Lineal se aprecian valores similares, entre $[10.5, 12]$ %; en RBF el error promedio es el mismo para todos los experimentos, un 12.06 %.

Tabla 1. Errores promedio obtenidos tras los experimentos de optimización y clasificación realizados.

SA-SVM		GA-SVM		PSO-SVM	
Kernel Lineal	RBF	Kernel Lineal	RBF	Kernel Lineal	RBF
0.102	0.1205	**0.0909**	0.128	0.1076	0.1206
0.102	0.1206	0.1095	0.1169	0.1206	0.1206
0.0965	0.1187	0.1113	0.1169	0.1057	0.1206
0.0946	0.115	0.1484	0.1243	0.1094	0.1206
0.0946	0.115	0.1169	0.1187	0.1206	0.1206
0.1039	0.1224	0.1169	0.1336	0.1169	0.1206
0.1039	0.1206	0.1169	0.1224	0.1169	0.1206
0.102	0.1206	0.1058	0.1206	0.1113	0.1206
0.102	0.1169	0.1076	0.1317	0.1113	0.1206

4 Conclusiones

Se presentó una comparativa de tres metaheurísticas para obtener valores óptimos en la clasificación de vinos chilenos según su denominación de origen, usando sus índices de concentración de metales. Se implementaron 77 SVM de salida binaria aplicadas en tres metaheurísticas, usando tanto Kernel Lineal como RBF. El entrenamiento se realizó mediante LOOCV. Los resultados se obtuvieron en base al error promedio para cada clasificación. Estos permiten obervar la eficiencia del Recocido Simulado, con porcentajes oscilando alrededor del 10 % al combinarlo con Kernel Lineal. Los errores alcanzados por Algoritmos Genéticos llegan casi al 15 %, y en Enjambre de Partículas alrededor del 12 %. Se observa la posible separabilidad lineal de los datos al obtener mejores resultados empleando Kernel Lineal en los tres algoritmos.

Bibliografía

1. Mandrile, L., Zeppa, G., Giovannozi, A., Rossi, A.: Controlling protected designation of origin of wine by raman spectroscopy. Food Chemistry **211** (2016) 260–267
2. Beltrán, N., Duarte-Mermoud, M., Bustos, M., Salah, S., Loyola, E., na Neira, A.P., Jalocha, J.: Feature extraction and classification of chilean wines. Journal of Food Engineering **75** (2006) 1–10
3. Villagra, E., Santos, L., Vaz, B.G., Eberlin, M., Laurie, F.: Varietal discrimination of chilean wines by direct injection mass spectrometry analysis combined with multivariate statistics. Food Chemistry **131** (2012) 692–697

4. Laurie, F., Villagra, E., Tapia, J., Sarkis, J., Hortellini, M.: Analysis of major metallic elements in chilean wines by atomic absorption spectroscopy. Ciencia e Investigación Agraria **37** (2010) 77–85
5. Fabani, M., Arrúa, R., Vásquez, F., Díaz, M., Baroni, M., Wunderlin, D.: Evaluation of elemental profile coupled to chemometrics to assess the geographical origin of argentinean wines. Food Chemistry **119** (2010) 372–379
6. Martelo-Vidal, M., Domínguez-Agis, F., Vásquez, M.: Ultraviolet/visible/near-infrared spectral analysis and chemometric tools for the discrimination of wines between subzones inside a controlled designation of origin: a case study of rías baixas. Australian Journal of Grape and Wine Research **19** (2013) 62–67
7. Martelo-Vidal, M., Vásquez, M.: Analysis of major metallic elements in chilean wines by atomic absorption spectroscopy. Ciência e Técnica Vitivinícola **29** (2014) 35–43
8. Liang, N., Liu, Y., Wang, L., Wang, P., Wang, J., Han, S.: Differentiation of geographical origins for cabernet sauvignon wines. Technical report, Agilent Technologies Inc. (2016)
9. Acevedo, F., Jiménez, J., Maldonado, S., Domínguez, E., Narváez, A.: Classification of wines produced in specific regions by uv-visible spectroscopy combined with support vector machines. Journal of Agricultural and Food Chemistry **55** (2007) 6842–9
10. Gaspar, P., Carbonell, J., Oliveira, J.: On the parameter optimization of support vector machines for binary classification. Journal of Integrative Bioinformatics **9** (2012) 201
11. Damaševičius, R.: Optimization of svm parameters for recognition of regulatory dns sequences. TOP **18** (2010) 339–353
12. Mantovani, R., Rossi, A., Vanschoren, J., B, B., de Carvalho, A.: Effectiveness of random search in svm hyper-parameter tuning. IEEE Proceedings of the 2015 International Joint Conference on Neural Networks, Killarney, Ireland (2015)
13. Diosan, L., Rogozan, A., Pecuchet, J.: Improving classification performance of support vector machine by genetically optimising kernel shape and hyper-parameters. Applied Intelligence **36** (2012) 280–294
14. Chen, H., Yang, B., Wang, S., Wang, G., Liu, D., Li, H., Liu, W.: Towards an optimal support vector machine classifier using a parallel particle swarm optimization strategy. Applied Mathematics and Computation **239** (2014) 180–197
15. Bao, Y., Hu, Z., Xiong, T.: A pso and pattern search based memetic algorithm for svms parameters optimization. Neurocomputing **117** (2013) 98–106
16. Li, S., Tan, M.: Tuning svm parameters by using a hybrid clpso-bfgs algorithm. Neurocomputing **73** (2010)
17. Alwan, H., Ku-Mahamud, K.: Incremental continous ant colony optimization technique for support vector machine model selection problem. 3rd International Conference on Applied Mathematics and Informatics (AMATHI '12), Montreux, Switzerland (2012)
18. Huang, C., Wang, C.: A ga-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications **31** (2006) 231–240
19. Sartakhti, J., Zangooei, M., Mozafari, K.: Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (svm-sa). Computer Methods and Programs in Biomedicine **108** (2012) 570–579
20. Lin, S., Lee, Z., Chen, S., Tseng, T.: Parameter determination of support vector machine and feature selection using simulated annealing approach. Applied Soft Computing **8** (2008) 1505–1512