Aplicación del modelo hibrido k-nearest neighbors-Support Vector Machine para la predicción del riesgo psicosocial en docentes de colegios públicos colombianos

Rodolfo Mosquera Navarro, Ph.D Student[1], Omar Danilo Castrillón Gómez, Ph.D [2], and Liliana Parra Osorio, Ph.D [3]
[1,2]Universidad Nacional de Colombia - Sede Manizales - Facultad de Ingeniería y Arquitectura – Departamento de Ingeniería Industrial – Grupo de Innovación y Desarrollo Tecnológico - Bloque Q Campus La Nubia, Manizales, 170001 – Colombia.
[3]Universidad Libre, Seccional Bogotá, Docente Investigador, Centro de Investigaciones Socio jurídicas, Colombia.
(E-mail: rmosqueran@unal.edu.co, odcastrillong@unal.edu.co)

Abstract*– Along with the increase of multiple task at the work, the prediction of risk levels becomes very significant. If the psychosocial risk levels is high, the work force in the companies will obtain a decrease in its productivity related to the occupational health and safety. Support vector machine is one of the most popular algorithms for the identification of the psychosocial risk levels. However, there is a disadvantage that the more close to the optimal hyperplane, the greater possibility of error label of the data. In this paper, we employ the hybrid model k-nearest neighbors-Support Vector Machines proposed by Zhou, this algorithm improve the classification and prediction accuracy of Support Vector Machines. This way fully combines the advantages of k-nearest neighbors and Support Vector Machines algorithms. The school-teachers psychosocial risk level datasets are used in our experiments. The experimental results shows that the hybrid k-nearest neighbors-Support Vector Machines model is a promising approach for the prediction of the psychosocial risk level in public school teachers with 86,66% accuracy.

Keywords—KNN; SVM; Psychosocial risk; School teachers; Hybrid model.

I. INTRODUCCION

Con el rápido desarrollo de los modelos de medición de la productividad, el riesgo psicosocial en el ambiente ocupacional de las escuelas públicas está aumentando y las pérdidas por el ausentismo aumentan. La evaluación de los riesgos psicosociales en docentes de colegios públicos se ha vuelto muy importante. Un buen modelo de predicción puede ayudar a los rectores a tomar decisiones correctas para disminuir el riesgo latente.

En la actualidad, hay tantos métodos mejorados basados en SVM [1], como PSO-SVM [2], [3], [4], GA-SVM [5], [6], GC-ABC [7] y FCM [8], [9], [10], la precisión y la eficiencia de la clasificación de los datos se ha incrementado, lo que depende principalmente de los parámetros Support Vector Machine que optimizan las propiedades en la selección de datos y reducen el número de vectores de soporte de SVM. En este documento, describimos un modelo híbrido de k-nearest neighbors-Support Vector Machine (KNN-SVM), que se basa en tratar con el conjunto de datos de predicción para lograr aumentar toda la precisión de la predicción de las máquinas de soporte vectorial (SVM).

El resto de este documento está organizado de la siguiente manera: la sección II describe el concepto básico; la sección III da el detalle del modelo híbrido k-nearest neighbors-Support Vector Machine (KNN-SVM), en la sección IV, demostramos el experimento del conjunto de datos y el análisis de los resultados; la sección V da una conclusión y propone futuras investigaciones.

II. CONCEPTOS DE K-NEAREST NEIGHBORS Y SUPPORT VECTOR MACHINE

A. *K Vecinos más Cercanos (KNN)*

K vecinos más cercano (K nearest neighbor) [16], [17], [18] es un método de aprendizaje supervisado perezoso, el cual está basado en el estudio de las etiquetas de datos de test que se compara con las etiquetas similares del grupo de entrenamiento.

Los datos de entrenamiento se pueden describir mediante la propiedad de la distancia. Supongamos que cada dato representa un punto en un espacio dimensional *d*, cada punto dentro del conjunto de entrenamiento puede ser almacenado dentro de un espacio d dimensional de patrones. Cuando se asigna una tupla, el algoritmo K nearest neighbor (KNN) puede realizar la búsqueda en el espacio del modelo y encontrar las tuplas mas cercanas *K* dentro del conjunto de datos de entrenamiento. Estos puntos *K* son los vecinos más cercano *K* de un punto desconocido.

La proximidad de las tupla con patrones similares se puede calcular a través de la utilización de la distancia Euclidea. La distancia euclidea entre dos puntos $X_1 = \{x_{11},x_{12},\ldots,x_{1d}\}$ y $X_2 = \{x_{21},x_{22},\ldots,x_{2d}\}$:

$$Dist(X_1,X_2) = \sqrt{\sum_{i=1}^{d}(x_{1i} - x_{2i})^2} \tag{9}$$

Para la clasificación K nearest neighbor (KNN), las etiquetas desconocidas de las tuplas deben ser asignadas a las etiquetas del K vecino más cercano. En otras palabras, las etiquetas de los puntos de predicción deberían estar asignados a las etiquetas de los vecinos más cercanos *K*.

B. Máquinas de Soporte Vectorial (SVM)

Basado en el principio de minimización del riesgo de estructural [11], SVM puede encontrar los datos (vector de soporte) que tienen una mejor capacidad para distinguir y construir el hiperplano óptimo que pueden ser los dos tipos de segmentación. Supongamos un conjunto de entrenamiento $X = (x_1; x_2; \ldots, x_n)$ y sus correspondientes etiquetas $Y = (y_1; y_2; \ldots, y_n,$ el cual puede expresarse como: $(xi; yi); xi \in Rd; yi \in (+1; -1); i \in (1; 2; \ldots, n)$ donde, d es el número de dimensión del espacio de entrada, n es el número de muestras.

En la condición de separación lineal, hay un hiperplano óptimo para distinguir las dos clases de muestras. Ver Fig. 1. Este hiperplano puede ser descrito como en la ecuación 1:

$$(w^* x_i) + b = 0, \quad (1)$$

Los datos se clasificarán acorde a las siguientes ecuaciones:

$$(w * x_i) + b \geq 0, y_i = +1, \quad (2)$$

$$(w * x_i) + b < 0, y_i = -1, \quad (3)$$

Donde, w es la dirección normal del hiperplano óptimo.

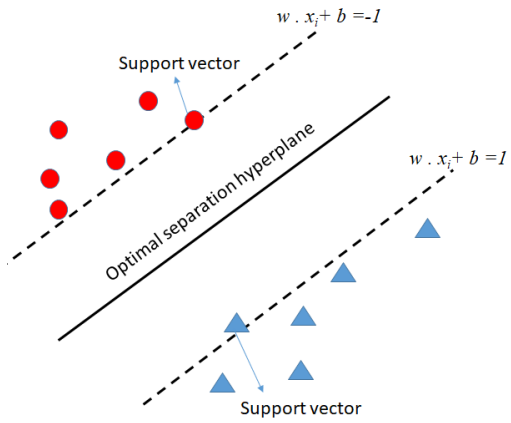

Fig. 1. Hiperplano óptimo de separación.

Si los datos de entrenamiento se clasifican correctamente con el menor error, y la distancia entre los datos del hiperplano más cercano y el hiperplano más lejano, podemos considerar al hiperplano como el hiperplano óptimo. En el caso de la separación lineal, una solución para obtener el hiperplano óptimo debería tratarse como una solución de un problema de programación cuadrática. Para las muestras de entrenamiento dadas, lo importante es encontrar los pesos óptimos w y el sesgo b, que minimiza los pesos de la función de costo de la siguiente manera:

$$\min S(w) = \frac{1}{2}\|w\|^2, \quad (4)$$

Sujeto a,

$$y_i[(w * x_i) + b] - 1 \geq 0, i = 1,2, \ldots l, \quad (5)$$

Donde, la función de optimización $S(w)$ es cuadrática y la restricción es lineal, lo que conlleva a una tradicional ecuación cuadrática de programación. Entonces:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i [y_{i,}[(w * x_i) + b] - 1] \quad (6)$$

Donde, $\alpha_i \geq 0 (i=1,2,\ldots l)$ es un multiplicador langraniano y el valor extremo de L es un punto a optimizar, la derivada de L puede hallar la solución óptima. Entonces, la función final sería:

$$plabel = sign(w^* * x + b), \quad (7)$$

Donde, w^* y b son la solución óptima del modelo. Sin embargo, los datos aquí presentados no son todos linealmente separables.

En caso de que los datos no sean linealmente separables, se debe mapear en el espacio de búsqueda de alta dimensionalidad como lograr una adecuada separación de los datos. La función kernel [12] que acompaña a la máquina de soporte vectorial debería ser capaz de mapear este proceso. La función kernel $Mr(x_i,x_j)$ satisface el teorema de mercer [13], [14], [15], y el producto interno con los datos de ejemplo podría calcularse con la función kernel. Actualmente, los métodos kernel mas usados son la función lineal, la función polinomial y la función radial.

La función final puede ser descrita como:

$$plabel = sing(\sum_{i=1}^{m} w_i Mr(x_i, x) + b), \quad (8)$$

Donde, m es en número de vectores de soporte y w_i es el peso del vector.

III. MATERIALES Y MÉTODOS

Utilizamos el modelo hibrido propuesto por Zhou et al., 2013 para la predicción del riesgo financiero adaptado a la predicción del riesgo psicosocial, para evaluar la calidad del algoritmo respecto a la eficiencia en la precisión de clasificación de los datos de riesgo psicosocial en docentes de colegios públicos en Colombia. Se escogió éste algoritmo hibrido como parte de un proceso de evaluación de diferentes alternativas para lograr obtener resultados que permitan dar una aproximación sobre el desempeño en el tipo de datos a clasificar.

De acuerdo al problema mencionado anteriormente, la probabilidad de que el test de datos cercano al hiperplano

optimo este mal clasificado es alta. Para ello el modelo hibrido pretende calcular los promedios de las distancias entre todos los vectores de soporte y el hiperplano y computa las distancias entre cada dato del test de datos y el hiperplano. Entonces, lo que se pretende con este cálculo básico de distancias, es dividir el test de datos en dos partes. Con base en la ecuación (5) y (6), la fórmula de la distancia puede ser expresada así (10):

$$d = \frac{|\sum_{i=1}^{m} w_i K(x_i, x) + b|}{\|w\|} \quad (10)$$

Donde el valor de ||w|| es constante a la vez que calcula la distancia, la distancia relativa se puede definir de la siguiente manera: (11):

$$d = \left| \sum_{i=1}^{m} w_i K(x_i, x) + b \right| \quad (11)$$

Dónde las distancias del test de datos se clasifican lo mejor posible cercano al hiperplano óptimo buscando obtener el menor error posible de clasificación (Fig. 2).

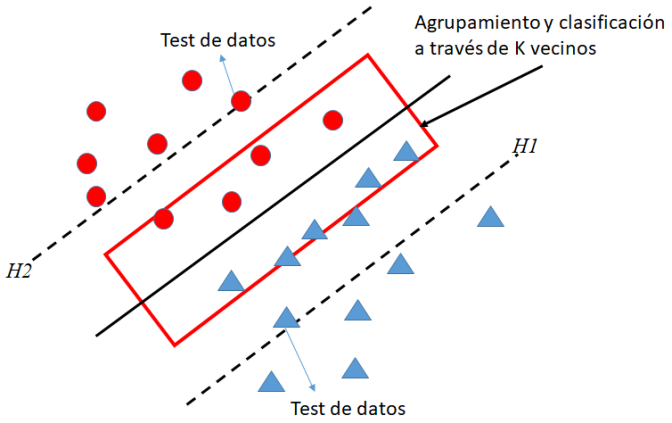

Fig. 2. Aproximación del modelo híbrido.

IV. EXPERIMENTOS

A. *Paso 1: Dataset*

Para modelar el comportamiento del riesgo psicosocial usamos los datos de riesgo psicosocial recolectados en docentes de colegios públicos colombianos. El conjunto de datos incluye 495 instancias y consta de 31 características. Se utilizó el 80% de los datos de entrenamiento y el 20% de los datos para la validación.

Tabla I
Información del Dataset

Nombre	Total Datos	Datos Entrenamiento(80%)	Datos Validación(20%)
Psicosocial	495	396	99

B. *Paso 2: Resultados Experimento*

Se ejecutó el algoritmo en Matlab V9.4 con la clasificación Support Vector Machine, se utilizó el kernel polinomial el cual utiliza la función y se ejecuta la distancia relativa con la siguiente ecuación:

$$"d = (W\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j (x_i * x_j)" \quad (12)$$

Los parámetros de la función kernel y del clasificador SVM son dados por defecto.

Tabla II
Error de clasificación

Nombre	Error de Clasificación	Mid	Porcentaje
Psicosocial	15	10	66,66%

La columna de error de clasificación representa el número total de muestras de prueba clasificadas erróneamente por SVM.

La columna del medio muestra el número de muestras de prueba cerca del hiperplano mal clasificado por SVM. La columna de porcentaje representa el porcentaje del número medio y el número de clasificación errónea. De la tabla anterior, los valores porcentuales son casi mayores que 66,66%. Por lo tanto, cuando se emplea SVM para predecir, la mayoría de los datos clasificados erróneamente se encuentran entre dos bordes y el hiperplano óptimo.

Tabla III
Nivel de Precisión

Nombre	KNN+SVM	SVM	KNN
Psicosocial	86,66%	84,75%	84,66%

Del resultado anterior, podemos llegar a la conclusión que el modelo híbrido de k-nearest neighbors-Support Vector Machine (KNN-SVM) tiene una mayor precisión que el algoritmo original de Support Vector Machine y k-nearest neighbors en la identificación y predicción del grado de riesgo psicosocial en docentes de colegios públicos de Colombia (Fig.3).

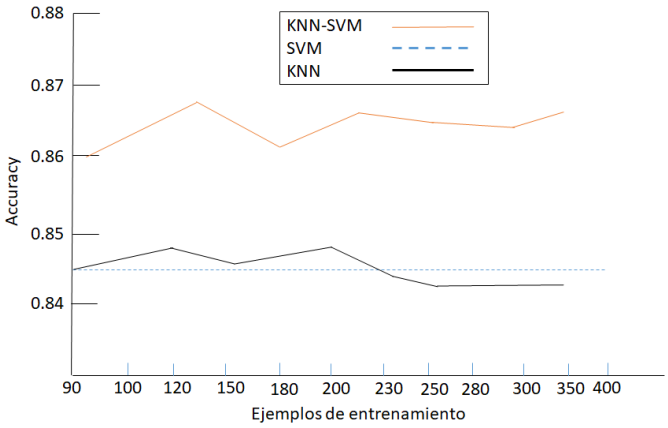

Fig. 3. Porcentaje de predicción de los modelos.

Con el fin de probar la influencia del modelo híbrido junto con el crecimiento de las muestras de entrenamiento, realizamos otro conjunto de experimentos. La cantidad de muestras de entrenamiento crecerá en regularidad y se agregará treinta o más en cada experimento. De las dos cifras anteriores, se puede descubrir que la precisión promedio del modelo híbrido es mayor que las otras, y con el aumento del número de muestras de entrenamiento, la precisión predicha del modelo híbrido mantiene una tendencia al alza, mientras que las otras se mantiene estable o mantiene una tendencia a la baja para el conjunto de datos riesgo psicosocial en docentes de colegios públicos de Colombia.

V. CONCLUSIONES

El diagnóstico de riesgo psicosocial asociado a las condiciones laborales ha sido de especial interés en los últimos años desde la perspectiva de la predicción del riesgo a través de técnicas inteligentes [19], [20], [21]. En este documento, adaptamos, aplicamos y presentamos el desempeño de un algoritmo Support Vector Machine (SVM) mejorado basado en k-nearest neighbors (KNN) (Zhou et al., 2013) y lo aplicamos en la identificación y predicción del riesgo psicosocial en docentes de colegios públicos de Colombia en cinco municipios de un área metropolitana. El algoritmo implementado combina el SVM y el KNN logrando ventajas significativas con respecto a los algoritmos individuales y mejora la clasificación. Se logró obtener un nivel de clasificación del 86.66% con una tasa de error del 13.34% con respecto a un 84.75% de máquinas de soporte vectorial y un 84.66% de k-vecinos más cercano.

Por lo tanto, el modelo híbrido puede mejorar la precisión de la predicción del SVM y KNN originales en cierto porcentaje. Sin embargo, el tiempo de procesamiento en el modelo conlleva mucho tiempo, lo que deja una línea abierta de investigación para futuros trabajos.

AGRADECIMIENTOS

Agradecemos a la “Convocatoria Nacional para el Apoyo al Desarrollo de Tesis de Posgrado o de Trabajos Finales de Especialidades en el área de la Salud de la Universidad Nacional de Colombia 2017-2018”, Resolución 21 de 2017 de la oficina del vice-rector de investigación (21 de diciembre de 2017) por la selección de la propuesta de investigación con el número de identificación 40976. También agradecemos a la Universidad Nacional de Colombia, sede Manizales por el apoyo en la investigación llevada a cabo por el autor del presente artículo como parte del proyecto de tesis doctoral en Ingeniería, industria y organizaciones como parte parcial de los resultados mostrados aquí.

REFERENCES

[1] Cortes, Corinna, and Vladimir Vapnik. ”Support-vector networks.” Machine learning 20.3 (1995): 273-297.
Tu, Chung-Jui, et al. ”Feature selection using PSO-SVM.” IAENG International journal of computer science 33.1 (2007):111-116.
[2] Ling, Yun, Qiu-yan Cao, and Hua Zhang. ”Application of the PSO-SVM model for Credit Scoring.” Computational Intelligence and Security (CIS), 2011 Seventh International Conference on. IEEE, 2011.
[3] Huang, Cheng-Lung, and Jian-Fan Dun. ”A distributed PSOSVM hybrid system with feature selection and parameter optimization.” Applied Soft Computing 8.4 (2008): 1381-1391.
[4] Huerta, Edmundo Bonilla, Batrice Duval, and Jin-Kao Hao. ”A hybrid GA/SVM approach for gene selection and classification of microarray data.”Applications of Evolutionary Computing. Springer Berlin Heidelberg, 2006. 34-44.
[5] Yu, Enzhe, and Sungzoon Cho. ”GA-SVM wrapper approach for feature subset selection in keystroke dynamics identity verification.” Neural Networks, 2003. Proceedings of the International Joint Conference on. Vol. 3. IEEE, 2003.
[6] Li, Lijie. ”A Novel Algorithm for Kernel Optimization of Support Vector Machine.” Advances in Swarm Intelligence. Springer Berlin Heidelberg, 2013. 98-105.
[7] Sivakumar, S., and C. Chandrasekar. ”Lung Nodule Detection Using Fuzzy Clustering and Support Vector Machines.” International Journal of Engineering and Technology 5 (2013).
[8] Zhang, Yong, et al. ”A novel fuzzy compensation multi-class support vector machine.” Applied Intelligence 27.1 (2007): 21-28.
[9] Tang, Yuchun, Yan-Qing Zhang, and Zhen Huang. ”FCMSVM-RFE gene feature selection algorithm for leukemia classification from microarray gene expression data.” Fuzzy Systems, 2005. FUZZ’05. The 14th IEEE International Conference on. IEEE, 2005.
[10] Shawe-Taylor, John, et al. ”Structural risk minimization over data-dependent hierarchies.” Information Theory, IEEE Transactions on 44.5 (1998): 1926-1940.
[11] Cristianini, Nello, and John Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.
[12] Zhou, Shangming, and John Q. Gan. ”Mercer kernel, fuzzy cmeans algorithm, and prototypes of clusters.” Intelligent Data Engineering and Automated LearningIDEAL 2004. Springer Berlin Heidelberg, 2004. 613-618.
[13] Minh, Ha Quang, Partha Niyogi, and Yuan Yao. ”Mercers theorem, feature maps, and smoothing.” Learning theory. Springer Berlin Heidelberg, 2006. 154-168.
[14] Lyu, Siwei. ”Mercer kernels for object recognition with local features.” Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 2. IEEE, 2005.
[15] Soucy, Pascal, and Guy W. Mineau. ”A simple KNN algorithm for text categorization.” Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001.
[16] Zhang, Min-Ling, and Zhi-Hua Zhou. ”ML-KNN: A lazy learning approach to multi-label learning.” Pattern Recognition 40.7 (2007): 2038-2048.

[17]Soucy, Pascal, and Guy W. Mineau. ”A simple KNN algorithm for text categorization.” Data Mining, 2001. ICDM 2001,Proceedings IEEE International Conference on. IEEE, 2001.
[18]Mosquera, Rodolfo, Castrillon, OD, and Parra Osorio, L.” Metodología para la predicción del grado de riesgo psicosocial en docentes de colegios colombianos utilizando técnicas de minería de datos.” Inf.tecnol 27.6 (2016): 259-272.
[19]Mosquera, Rodolfo, Castrillon, OD, and Parra Osorio, L.” Predicción de riesgos psicosociales en docentes de colegios publicos colombianos utilizando técnicas de inteligencia artificial.” Inf.tecnol 29.4 (2018): 267-280.
[20]Mosquera, Rodolfo, Castrillon, OD, and Parra Osorio, L.” Máquinas de soporte vectorial, clasificador naive bayes y algoritmos genéticos para la predicción de riesgos psicosociales en docentes de colegios públicos colombianos.” Inf.tecnol 29.6 (2018): 153-162.
[21] Lotfan, S., Shahyad, S., Khosrowabadi, R., Mohammadi, A., & Hatef, B. “Support vector machine classification of brain states exposed to social stress test using EEG-based brain network measures.” Biocybernetics and Biomedical Engineering, 39.1, (2019): 199-213.