TEMA:

“LA CALIDAD DE LA CONEXIÓN INALÁMBRICA Y SU RELACIÓN CON LAS CONDICIONES METEOROLÓGICAS”

AUTORES: Ing. Rodolfo Najarro Quintero MSc. - Lcdo. Amilkar Yudier Puris Cáceres. PhD.

RESUMEN

El desarrollo de este trabajo de investigación se realiza con el objetivo de encontrar la relación que existe entre las condiciones meteorológicas y la conexión inalámbrica en base terrena. Los datos son suministrados por un centro de meteorología de la zona y una empresa de telecomunicación que opera en el mismo lugar.

Se estudia directamente los modelos basado en lógica difusa debido a que la fácil interpretación de las reglas y manejo de los datos. Para este proceso se utiliza el software Weka que ofrece herramientas para clasificación y el pre procesamiento de datos, así como el software Keel para la aplicación de diferentes clasificadores basados en reglas difusas.

Se aplicaron nueve clasificadores difusos donde el Furia-C fue el que mejores resultados obtuvo en cuanto a cantidad de reglas y calidad de clasificación. En este escenario se realizó una etapa de pre procesamiento donde fueron utilizado algunas técnicas para mejorar la información. Algunas de las reglar obtenidas corroboran la influencia que tiene la lluvia fuerte sobre la pérdida de la señal, pero aparecen otras relaciones que incorporan nuevos conocimientos en el área, como por ejemplo el punto de rocío y la humedad relativa media.

Palabras claves: Minería de datos, estación meteorológica, calidad de conexión, predicción de variables.

INTRODUCCIÓN

En la actualidad el desarrollo de las telecomunicaciones está orientado al uso de las tecnologías digitales, reduciendo en gran medida las transmisiones analógicas. Su evolución se ha visto notablemente en los sistemas de radiocomunicación por satélite para radiodifusión digital directa de TV y audio, comunicaciones móviles de banda estrecha, y novedosos proyectos de comunicaciones fijas de banda ancha, utilizando tanto satélites geoestacionarios como no-geoestacionarios.

Uno de los problemas más notorios dentro de las redes de comunicación inalámbrica representa la perdida de la señal debido entre otros factores a las condiciones meteorológicas. La lluvia, la nubosidad, la temperatura, el viento y la humedad son algunas de las variables que influyen en la perdida de paquetes de datos transmitidos.

Un mecanismo que permita predecir con anterioridad la perdida de la señal inalámbrica en dependencia de las condiciones meteorológicas, ayudaría a tomar decisiones alternativas para no perder información importante en los canales de comunicación.

Entre los modelos de predicción más utilizados para la ayuda de la toma de decisiones se encuentra el razonamiento basado en reglas difusas. La lógica difusa a diferencia de la lógica clásica o de Aristóteles procura crear aproximaciones matemáticas en la resolución de ciertos tipos de problemas donde existen datos imprecisos. Por otra parte, los sistemas basados en reglas difusas utilizan la lógica difusa para representar la incertidumbre de las variables para construir reglas lógicas que simulen un razonamiento inteligente a nivel de expertos en el área.

Este es el contexto donde se enmarca la presente investigación donde se propone la generación de reglas difusas a partir de datos reales climatológicos que ayuden a

predecir la perdida de la señal inalámbrica en beneficio de los servicios de las telecomunicaciones.

METODOLOGÍA

PROCEDIMIENTO DE LA INVESTIGACIÓN

La realización del trabajo de investigación se centralizó en el campus "Ing. Manuel Haz Álvarez" ubicado en el Km 1/2 de la vía a Santo Domingo, cantón Quevedo, provincia de Los Ríos. Una institución de educación superior con ardua experiencia en el campo agrícola y con nuevas carreras técnicas que surgen para el fortalecimiento de la misma, creada con el fin de aportar profesionales bien capacitados a la región.

La recolección de la información se realizó a través de entrevista a especialistas del INAMHI y de la empresa TELCONET donde caracterizaron los datos y brindaron toda la información necesaria para la elaboración de la investigación.

Descripción de la Información Obtenida.

Para realizar nuestra investigación e implementación de un modelo difuso basado en reglas para caracterizar la calidad de la señal inalámbrica en dependencia de las variables meteorológicas se realizó una investigación documental mediante la consulta libros de telecomunicaciones y meteorología, artículos científicos de diferentes autores, bases de conocimientos donde se encuentran estudios realizados sobre modelos difusos aplicados a otras ramas del conocimiento y la realización de consultas a expertos sobre el tema de reglas difusas y meteorología.

Se corrobora la consistencia de los datos meteorológicos a través de un análisis manual y automático de cada uno de ellos por ejemplo revisión de los rangos de cada variable que se ajusten a las condiciones reales del entorno local también se trabajó con relación a la

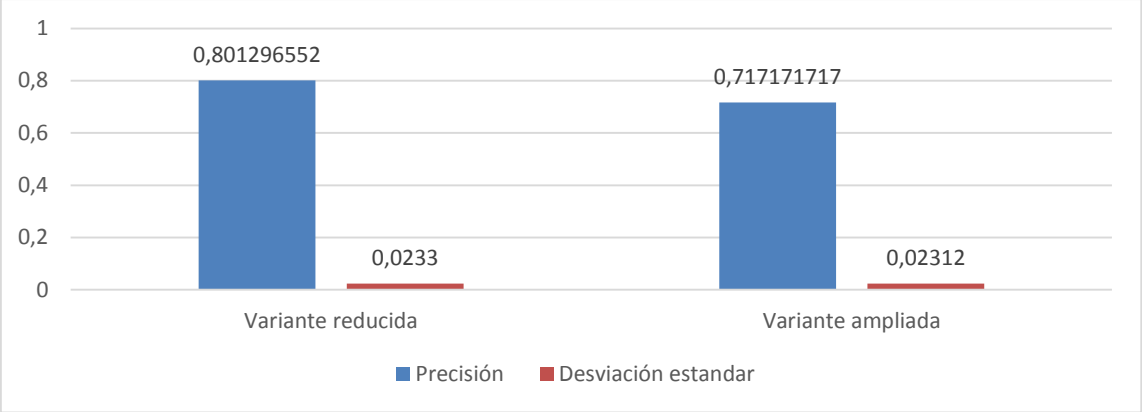
imputación de datos relacionados con los valores perdidos para casos en donde no se encontraban valores

ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS

La investigación se realizó utilizando un modelo difuso basado en reglas para caracterizar la pérdida o no de la señal inalámbrica en dependencia de diferentes variables meteorológicas, en la realización de los experimentos se aplicaron diferentes algoritmos, que a través de clasificadores se discretizaron las variables y seleccionaron atributos donde con los valores obtenidos se determinó cual presento la mejor alternativa en cada experimento realizado utilizando 3 indicadores como son la precisión, la desviación estándar y la cantidad de reglas.

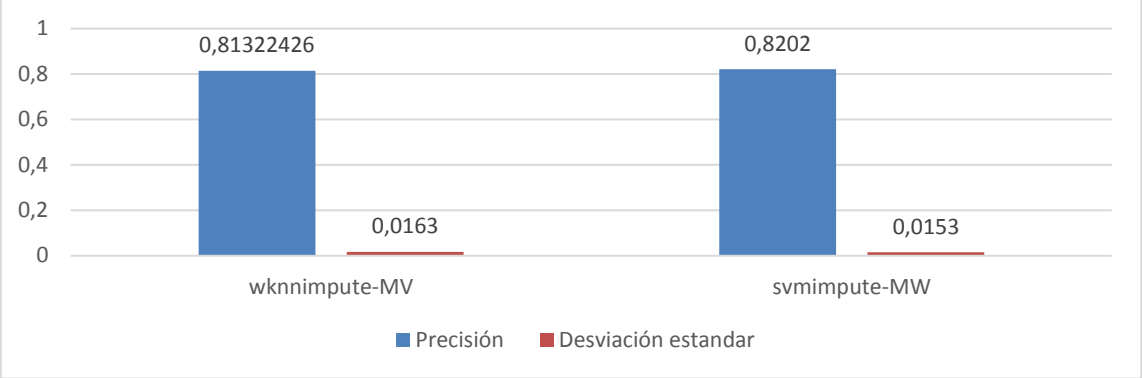
La calidad de los resultados obtenidos en el proceso de obtención de conocimientos no solo depende de los métodos de extracción de conocimiento sino también de cómo se haya conformado la base de conocimiento y todo el pre procesamiento desarrollado para obtener una base de conocimiento lo más sólida posible.

La base de conocimiento utilizada en la presente investigación contiene 2184 (3 mediciones diarias) mediciones meteorológicas que representan 728 días y de cada día se cuenta con el reporte de si se perdió o no la conexión inalámbrica, de manera que se realizaron dos posibles transformaciones de datos:Reducir el número de mediciones de cada variable a 728 a través del valor promedio de las 3 mediciones diarias (variante reducida) y Aumentar la cantidad de valores de la variable conexión copiando cada valor 3 veces uno por cada medición diaria (variante aumentada).

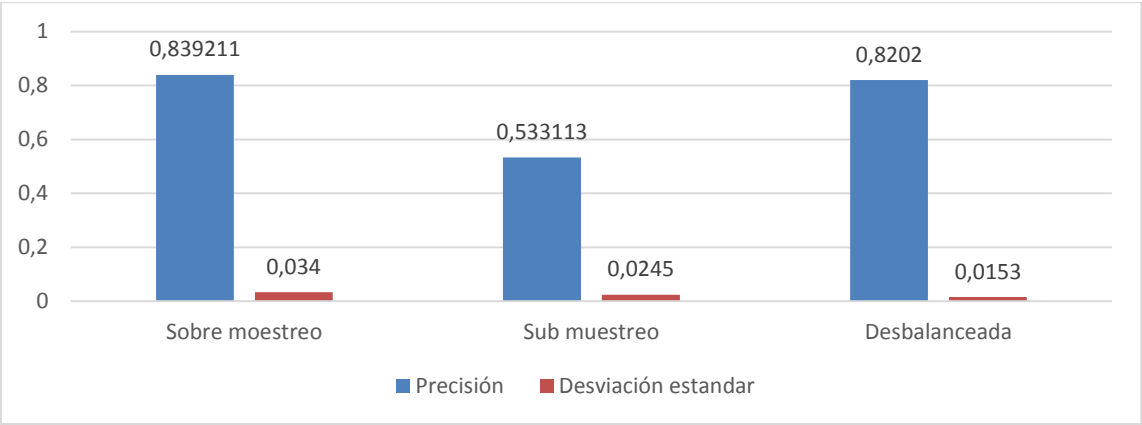


Como se puede apreciar en la figura la variante reducida obtiene mejores resultados en precisión y desviación estándar, debido a que las transformaciones realizadas en los datos para este caso se ajustan mucho más a la realidad. Por otra parte, el aumento de la información genera incertidumbre debido a la que el comportamiento en las 3 mediciones realizadas en el día puede ser totalmente diferente y tener el mismo comportamiento de la señal.

A continuación se aplicó dos de los algoritmos más utilizados en la literatura para la importación de datos Weighted Knn Imputation (wknnimpute-MV) (Troyanskaya O., 2001) y SVM Imputation (svmimpute-MW) (H.A.B. Feng, 2005) que aparecen que el software Keel. Para identificar cuál de ellos obtiene los mejores resultados, utilizando el clasificador difuso Furia-C, y en lo particular se puede observar que no existe diferencia significativa entre los resultados obtenidos por los dos algoritmos de imputación, aunque se presenta una leve mejora con el uso del imputador svmimpute-MW según se puede apreciar en la siguiente gráfica, el mismo que se utilizará para los siguientes experimentos.



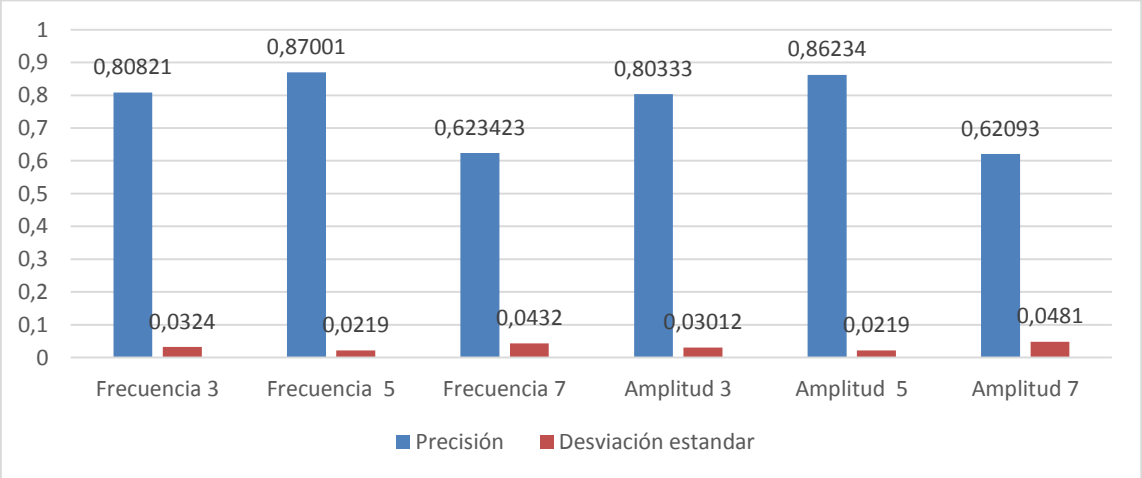
La presencia de clases no balanceadas en la base de conocimientos es uno de los problemas que se presentan más frecuentes, por lo que para eliminar el desbalance en la base de datos existen dos tipos de estrategias, el sobre muestreo y el sub muestreo, en esta investigación se utiliza el algoritmo SMOTE (Synthetic Minority Oversampling Technique), (Chawla, 2002) que sigue una estrategia de sobre muestreo donde las nuevas instancias se crean a través de la interpolación de los casos más cercanos a la clase minoritaria. Y el algoritmo SpreadSubsample (J.A. Olvera-Lopez, 2010) como una técnica de sub muestreo para eliminar casos de la clase mayoritaria.



Se puede apreciar como la técnica de sobre muestreo mejora la precisión del clasificador FURIA-C con respecto a no utilizar esta técnica, además se evidencia como la técnica de sub muestreo disminuye en gran medida la precisión. Esto se debe a que en la base de conocimientos se tiene poca cantidad de información y si se elimina parte de esta es muy difícil que se logre encontrar un modelo que responda de manera precisa a los

requerimientos del problema, por lo que se utilizará la base de datos resultante del proceso de sobre muestreo.

A continuación se sigue con el proceso de discretización de variables lo cual persigue obtener un nuevo conjunto de variables discretas a partir del conjunto original de datos continuos, producto a que la mayoría de los modelos de análisis de datos están preparados para trabajar con datos discretos además el trabajar con datos continuos aumenta considerablemente el número de reglas de los algoritmos basado en esta técnica, lo que atenta drásticamente con la eficiencia de los algoritmos y la muestra dio como resultado con el clasificador Furia-C aplicando dos de los algoritmos más utilizados en el estado del arte para la discretización de datos, Discretización por igual frecuencia y Discretización por igual amplitud aplicando en cada caso diferentes intervalos (3 ,5 ,7) para medir si la cantidad de intervalos influyen o no en la calidad del clasificado, que los mejores valores en cuanto a la precisión y la desviación estándar se alcanzan con el algoritmos de discretización por igual frecuencia para 5 intervalos, ya que estratos grandes y pequeños hacen carecer de interpretación al modelo.



Por otro lado, la selección de atributos o de variables es el proceso que elimina información redundante, haciendo posible que los algoritmos funcionen de forma más rápida y precisa, esta investigación consta de un grupo pequeño de atributos, específicamente 12 atributos predictores, no obstante se aplicaran algunos de los principales algoritmos para determinar

cuáles de los datos no está introduciendo ningún tipo de información relevante para el proceso de extracción de conocimiento con el objetivo de reducir la cantidad de reglas asociadas, por lo que se utilizó el software Weka donde se obtuvo como resultado que la mayoría de los algoritmos seleccionan un conjunto de 7 variables.

Algoritmo	Cantidad de atributos	Atributos
Best first	7	TM, TMAX, TOSC, HRM, HRMIN, PR, LL
SubsetSizeForwardSelection	7	TM, TMAX, TOSC, HRM, HRMIN, PR, LL
GeneticSearch	7	TM, TMAX, TOSC, HRM, HRMIN, PR, LL
GreedyStepwise	7	TM, TMAX, TOSC, HRM, HRMIN, PR, LL
ScatterSearchV1	7	TM, TMAX, TOSC, HRM, HRMIN, PR, LL
RandomeSearch	6	TM, TMAX, TOSC, HRMIN, PR, LL

A continuación, se realizó un breve estudio con el clasificador FURIA-C para identificar cuál de los subconjuntos obtenidos determina una mejor precisión y una menor cantidad de reglas y se seleccionara el algoritmo que mejor balance obtenga entre 3 indicadores fundamentales precisión, desviación estándar y cantidad de reglas e interpretación de las mismas lo cual se resume en la siguiente tabla,

Algoritmo	Precisión	Cantidad de reglas	Desviación estándar
FH-GBML-C	0,7492	11	0,0203
FURIA-C	0,8774	8	0,0199
GFS-GP-C	0,8009	14	0,0122
AdaBoost-C	0,7188	11	0,0328
GFSMaxLogitBoost-C	0,7144	13	0,0217
GFS-GPG-C	0,8276	14	0,0101
GFS-LogitBoost-C	0,8317	9	0,0239
SGERD-C	0,6618	7	0,0477
SLAVE-C	0,6509	15	0,0368

Según los resultados antes descritos se puede concluir que el algoritmo FURIA-C es el que mejor resultado de manera general obtiene, debido a que supera significativamente en precisión a los demás modelos y para los otros indicadores la diferencia con el mejor es mínima. A continuación se describe la base de reglas obtenidas por el algoritmo

Regla	Interpretación	Precisión
R1	TM mayor que 26,52 y LL mayor 104,72 perdida de señal	94 %
R2	PR mayor que 20,2 y HRM mayor que 95 and TOSC entre 3,6 y 6,4 perdida de señal	86 %
R3	TM entre 25,24 y 26,52 y TOSC menor que 3,6 y HRM entre 82,79 y 89,39 perdida de señal	63 %
R4	TOSC menor que 3,6 y LL entre 52,36 y 78,53 no hay perdida de señal	80 %
R5	PR mayor que 20,2 no hay perdida de señal	80 %
R6	HRMIN entre 38,93 y 56 no hay perdida de señal	79 %
R7	TM mayor que 26,52 no hay perdida de señal	76 %
R8	TM 23,96 y 25,24 TOSC menor 3,6 no hay perdida de señal	72 %

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

- Dentro de las 12 variables meteorológicas utilizadas en el proceso de extracción del conocimiento como son Lluvia, Temperatura Mínima, Temperatura Máxima, Temperatura Media, Humedad Relativa Máxima, Humedad Relativa Mínima, Humedad Relativa Media, Punto de Roció, Heliofanía, Oscilación Térmica, Tensión de Vapor de agua y Evaporación, solo 7 de ellas se relacionan directamente con la pérdida o no de la señal inalámbrica. Esto se obtuvo a través de proceso de selección de atributos donde se probaron diferentes algoritmos tipo filtro conjuntamente con el clasificador FURIA-C y los mejores resultados se obtuvieron con el subconjunto formado por Temperatura Media, Temperatura Máxima, Humedad Relativa Media, Humedad Relativa Mínima, Oscilación Térmica, Punto de Rocío y Lluvia.

- El modelo difuso obtenido responde a 8 reglas cuya interpretación se ajusta en parte a los conocimientos previos del fenómeno como es el caso de que cuando la temperatura media es alta y la lluvia es fuerte entonces la señal se pierde. No obstante, se obtuvieron otras reglas no tan fáciles de deducir por los expertos como es el caso de que cuando el punto de roció es alto, la humedad relativa es alta y la oscilación térmica es media la señal también se pierde.

- Se aplicaron un conjunto de 9 algoritmos basados en reglas difusas la mayoría responden a modelos evolutivos de búsqueda para encontrar cuál obtiene los mejores resultados. Identificando al clasificador FURIA-C como el que mejores resultados obtuvo en función de la precisión superando por más de, 4% al segundo mejor resultado. Para los otros indicadores sus resultados no difieren mucho de los modelos ganadores.

Recomendaciones

- Proponer a la empresa de telecomunicaciones que facilitó los datos para la investigación la implementación de un sistema con los resultados obtenidos para brindarle un mejor servicio a los clientes y garantizar que en realidad el % de servicio de calidad de señal que se anuncia se puede cumplir a través de esta herramienta.

- Involucrar en el estudio otro grupo de variables meteorológicas que no fueron tenidas en cuenta en esta investigación por falta de información, de manera que se puedan encontrar otros patrones importantes.

- Aplicar técnicas para descubrir variables ocultas que pueda ayudar a disminuir los niveles de imprecisión que atentan con la precisión del modelo obtenido.

- Utilizar otro grupo de clasificadores (arboles de decisión, modelos bayesianos, máquinas de soporte vectorial, entre otros) para aumentar los niveles de precisión de los clasificadores difusos utilizados en la presente investigación.

BIBLIOGRAFÍA

C., L. (2005). Determinación de Primeras Llegadas en Datos VSP y Check Shots usando Lógica Difusa. Universidad Simón Bolívar.

Carvalho, D., & Freitas, A. (2004). A hybrid decision tree/genetic algorithm method for data mining. *Information Sciences*, 13-35.

Chawla, N. (2002). SMOTE: Syntetic Minority . *Journal of Artificial Intelligence Research, 16*, 321-157.

Freeman, R. L. (2005). *Telecomunication System Engineering.*

G, D. J. (s.f.). *Curso Introductorio de Conjuntos y Sistemas Difusos.* Malaga, España: Universidad de Málaga.

González, M. (2006). *Utilización de Jats y lógica difusa en el análisis de dos perfiles de refracción sísmica en el área 75 de Bauxilum,*. Bolivar.

H.A.B. Feng, G. C. (2005). A SVM regression based approach to filling in Missing Values. *9th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES2005)* (págs. 581-587). Melbourne, Australia: Springer-Verlag.

Holton, J. (2004). *An Introduction to Dynamic Meteorology.* Amsterdam: Elsevier Academic Press.

Hüh, J., & E. H. (2009). FURIA: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery* , 293-319.

J.A. Olvera-Lopez, J. K.-O. (2010). A review of instance selection methods",. *Journal of Artificial Intelligence Review, 34*(2), 133-143.

J.Alcala-Fdez, A.Fernandez,J.Luego, J.Derrac. (2011). *Keel Datos Mineria Herramienta de Software.*

Jose Manuel Molina Lopez, J. G. (2006). *Tecnicas de Analisis de Datos.*

Liu, Y., Qin, Z., Shi, Z., & Chen, J. (2004). Rule Discovery with Particle Swarm Optimization. *Springer* , 291-296.

Molina, H. G. (2007). *Avances en Informatica y Sistemas Computacionales.* Mexico: Univ. J.Autonoma de Tabasco .

Morales, G. (2002). *Introducción a la lógica difusa.*

P.Reuterman, B. a. (2004). Proper: A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners. *17 th Australian Joint Conference on Artificial Intelligence Springer-Verlag* .

Rastogi, R., & Shim, K. (2000). A Decision Tree Classifier that Integrates Building and Pruning. *Data Mining and Knowledge Discovery*, 315-344.

Sierra, A. B. (2006). *Aprendizaje Automatico: conceptos basicos y avanzados.*

Troyanskaya O., C. M. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*, 520-525.

URUETA,L,VALDÉS,H. (2002). *La logica difusa como apoyo a la enseñanza.* Cartagena: Corporación Universitaria Rafael Núñez.

Zorrilla, M. E. (Febrero de 2003). Obtenido de personales.unican.es/zorrillm/MaterialOLD/redes.pd: www.google.com/search?