Adaptive Locally Weighted Random Forest

Akansha, Ruhani, Lucas

Motivation

Existing issues with random forest algorithm:

→All attributes are treated with equal weight

→Including attributes with low predictive power unnecessarily lowers accuracy

→Doesn't perform well on skewed data

Method

1. Create a random forest
2. Go through each test instance and retrieve k nearest neighbors
3. Calculate local feature weights using an attribute selection method (info gain, gain ratio, correlation, etc.)
4. Trees vote using aggregate weight value rather than one vote per tree
5. Label with largest weighted vote wins

Similar works remove attributes with low predictive power, change the likelihood of their selection, or weight classes based on distribution.

Dataset

Student Depression Dataset

- ~28k instances
- 17 attributes
- Binary class data (depressed or not)
- 2x as many depressed instances

Preprocessing

- Discretize class variable (weka NumericToNominal filter)
- Normalized numeric data using min-max normalization
- 70-30 train-test split

Dataset In-Depth

General Characteristics

ID, gender, age, city

Work/Academic Characteristics

Profession, CGPA (cumulative gpa), study satisfaction, job satisfaction, college degree, work/study hours

Stressors/Habits

Sleep duration, dietary habits, suicidal thoughts, financial stress, family history

CLASS VARIABLE: DEPRESSION

Relevant in today's mental health landscape, with high academic and outside stressors

Experiments

- Test our algorithm on multiple different attribute selection methods, compare to standard RF
  - Standard RF, KNN with Pearson Correlation, KNN with InfoGain, KNN with GainRatio
  - 100 trees per forest
    - log2 atts for each tree
    - 1/3 train set per sample
  - k = 100
- Conduct 3 trials of each
- Compare performance using various metrics (accuracy, precision, recall, specificity, confusion matrices)

Used python code to calculate attribute selection values for Pearson Correlation, InfoGain,and GainRatio.

Results

01

Average accuracy of the weighted forests of each attribute selection method

02

Standard RF

75.15%

03

Pearson Correlation

80.52%

04

InfoGain

80.61%

05

GainRatio

79.22%

06

used k-value of 100, this performed the best while still not being too computationally expensive

Results - Performance Metrics (Standard RF)

82.17%

Precision

TP / (TP + FP)

70.89%

Recall

TP / (TP + FN)

43.94%

Specificity

TN / (TN + FP)

Results - Performance Metrics (InfoGain)

82.07%

Precision

TP / (TP + FP)

78.45%

Recall

TP / (TP + FN)

64.75%

Specificity

TN / (TN + FP)

Shortcomings + Possible Improvements

- Specificity increased, but is still relatively low
  - Use smote for imbalanced class

- Determine best attribute selection method when classifying, instead of using one per run
- Optimize it further, since certain attribute selection methods (pearson correlation) take too long to run

References

Lin, R., He, X., Feng, J., Zalmout, N., Liang, Y., Xiong, L., & Dong, X. L. (2021). PAM: Understanding product images in cross product category attribute extraction [Paper presentation]. 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. https://doi.org/10.1145/3447548.3467164

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 10(1). https://doi.org/10.1186/1471-2105-10-213

Peker, M., Arslan, A., Şen, B., Çelebi, F. V., & But, A. (2015). A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+RF). 2015 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 1–8. https://doi.org/10.1109/INISTA.2015.7276737

Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. Journal of Biomedical Science and Engineering, 06(05), 551-560. https://doi.org/10.4236/jbise.2013.65070

Reif, D. M., Motsinger, A. A., McKinney, B. A., Crowe, J. E., & Moore, J. H. (2006). Feature selection using a random forests classifier for the integrated analysis of multiple data types. 2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, 1–8. https://doi.org/10.1109/CIBCB.2006.330987

R., U., & Ghosh, S. (2019). Maxima-turn-switching strategy of sensor-equipped UAVs for target localization in 2-D and 3-D environments. 2019 IEEE 15th International Conference on Control and Automation (ICCA), 610–615. https://doi.org/10.1109/ICCA.2019.8900007

Uduagbamen, P. K., Sanusi, M., Udom, O. B., Salami, O. F., Adebajo, A. D., & Alao, O. J. (2020). Metabolic acidosis in the surgical intensive care unit: Risk factors, clinical correlates and outcome. findings from a high dependency heart and vascular surgical center in nigeria. World Journal of Cardiovascular Surgery, 10(11), 226-241. https://doi.org/10.4236/wjcs.2020.1011025

THANK YOU

questions?

References

Lin, R., He, X., Feng, J., Zalmout, N., Liang, Y., Xiong, L., & Dong, X. L. (2021). PAM: Understanding product images in cross product category attribute extraction [Paper presentation]. 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. https://doi.org/10.1145/3447548.3467164

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 10(1). https://doi.org/10.1186/1471-2105-10-213

Peker, M., Arslan, A., Şen, B., Çelebi, F. V., & But, A. (2015). A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+RF). 2015 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 1–8. https://doi.org/10.1109/INISTA.2015.7276737

Nguyen, C., Wang, Y., & Nguyen, H. N. (2013). Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. Journal of Biomedical Science and Engineering, 06(05), 551-560. https://doi.org/10.4236/jbise.2013.65070

Reif, D. M., Motsinger, A. A., McKinney, B. A., Crowe, J. E., & Moore, J. H. (2006). Feature selection using a random forests classifier for the integrated analysis of multiple data types. 2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, 1–8. https://doi.org/10.1109/CIBCB.2006.330987

R., U., & Ghosh, S. (2019). Maxima-turn-switching strategy of sensor-equipped UAVs for target localization in 2-D and 3-D environments. 2019 IEEE 15th International Conference on Control and Automation (ICCA), 610–615. https://doi.org/10.1109/ICCA.2019.8900007

Uduagbamen, P. K., Sanusi, M., Udom, O. B., Salami, O. F., Adebajo, A. D., & Alao, O. J. (2020). Metabolic acidosis in the surgical intensive care unit: Risk factors, clinical correlates and outcome. findings from a high dependency heart and vascular surgical center in nigeria. World Journal of Cardiovascular Surgery, 10(11), 226-241. https://doi.org/10.4236/wjcs.2020.1011025