Analysis of Decision Tree and Smooth Support Vector Machine Methods on Data Mining

Natalina Br Sitepu\(^1\), Sawaluddin\(^1\), M Zarlis\(^1\), Syahril Efendi\(^1\), Hanna Willa Dhany\(^1\)

\(^1\) Informatic Engineering Study Program Magister, Faculty of Computer Science and Information Technology Universitas Sumatra Utara, Medan - Indonesia

* natalinasitepu89@gmail.com

Abstract. This research is about smooth support vector machine (SSVM) and Decision Tree in data mining. Many researchers conduct and develop methods to improve the accuracy and classification of data on good results. This research was conducted by conducting an experiment on STMIK Neumann Medan student data. In this study, it was concluded that Decision Tree performance is better than SSVM, Decision Tree gets very good results that promise to help find the best students to get scholarships. This study is better than SSVM. The training process has a difference of 11.04% and the testing process is 10.08% with each Accuracy.

1. Introducing

Information Technology is very influential in every life management, from its use in education to industry to make a management decision. With the development of information technology that has developed this has a positive impact on human life. Information technology is very valuable, because it provides benefits both directly and indirectly. So that humans become more productive in making changes and enhancing knowledge.

Data classification is a category of changes that are in accordance with the rules. In this classification aims to change the structure of the sample object. Classification algorithms are made from training sets and build models and models used to classify new objects. The Decision Tree evaluates the power of classification by performance analysis and the results of its analysis. Classification of data objects based on objects that have been specified in a data. there are many classification algorithms but the Decision tree is most commonly used. Decision Tree Algorithm is one of the most important classification measures in data mining. Classification is one type of grouping, which is a flowchart like a tree structure, where each internal node shows the test on each attribute, each branch represents the results of the test, and each leaf node represents the class The model for classifying a note to find the leaf root path to measure the test attributes and leaf attributes is the result of the classification used by the Decision Tree.

SVM utilizes optimization with quadratic programming, so for high-dimensional data and large amounts of data SVM becomes less efficient. Therefore, a smoothing technique is developed that replaces the SVM function with an integral of the sigmoid neural network function, hereinafter known as Smooth Support Vector Machine (SSVM). A machine-based learning method that is popular to
implement because it has high performance and can be widely applied for classification and estimation is a function of the Support vector machine.

In this study, two methods are used, namely Decision Tree and Smooth Support Vector Machine in finding students who deserve scholarships. Scholarships are money allowances has given to students or students who excel as tuition assistance to encourage other students to be more successful as well. From the background of the problems that have been described previously, the authors take the formulation of the problem that the need for a classification process in finding students who have the opportunity to get scholarships by using the Decision Tree and Smooth Support Vector Machine methods in order to produce more precise and optimal results.

2. Methodology
The classification process in this study is to find students who have the opportunity to get scholarships by using the Decision Tree and Smooth Support Vector Machine methods to produce more precise and optimal results.

The design of this study can be seen as follows:

![Research Flowchart](image)

**Figure 1. Research Flowchart**

2.1. Smooth Support Vector Machine
Image grouping point mn-dimensional real space \( \mathbb{R}^n \) that is owned by Smooth Support Vector Machine, which is found in \( m \times n \) matrix \( A \), on membership every point \( A_i \) in class 1 or -1 which is basically determined by the diagonal \( M_X \) on the matrix \( D \) with a minus along the diagonal. In this case SSVM with a linear kernel \( AA \) [30.12] at \( V > 0 \):

\[
\min_{(w,y,y) \in \mathbb{R}^{n+1+m}} \frac{1}{2} w^T w
\]

(1)

\[
\min_{(w,y,y) \in \mathbb{R}^{n+1+m}} \frac{1}{2} w^T w, \text{ s.t.} D(Aw - ey) + y \geq e
\]

(2)

\[y \geq 0\]
\[ x'w - \gamma = +1 \]
\[ x'w - \gamma = -1, \]

Y determines the relative original direction. When 2 classes are linearly separated between Class 1 and point -1, the variable \( y = 0 \).

\[ x'w = \gamma, \]

If the class is inseparably linear, they bind "soft margin" determined by non-negative variables \( y \), namely:

\[ x'w - \gamma + y_i \geq +1, \text{ for } x' = A_i \text{ and } D_{ii} = +1, \]
\[ x'w - \gamma - y_i \leq -1, \text{ for } x' = A_i \text{ and } D_{ii} = -1. \]

3. Results and Discussion

According to the discussion discussed, I try to research a study of student data sets. This research was carried out using a computer, which uses 1.66 GHz Intel atoms core™ 2 duo CPU processor. The computer runs using the Windows 7 operating system, by installing Rapid Miner 8.1. Classification of accuracy generated by SSVM can be seen in table 1 and the accuracy classification produced by the decision can be seen in the following table:

| SCORE | GRADE |
|-------|-------|
| 89-100 | A     |
| 75-88  | B     |
| 60-74  | C     |
| 40-59  | D     |
| 10-39  | E     |

Student academic achievement for each student is determined by the value scale table value scale. In Assessment of Student Attitudes Is:

GOOD : 5
MEDIUM : 3
BAD : 1

| D      | Name                      | GPA | Presence | Attitude |
|--------|---------------------------|-----|----------|----------|
| 152010001 | Ady Syah Putra Sitepu    | 3.14 | 3.24    | 3.12     | 80%  | B    |
| 152010002 | Megasasty Indah Lestari Gulo | 3.64 | 3.70    | 3.55     | 100% | B    |
| 152010004 | Febrian Hikma Risky S.   | 3.00 | 3.12    | 3.18     | 75%  | B    |
D | Name | GPA | Presence | Atitude |
|---|---|---|---|---|
| 152010005 | Angga Bremana | 3.24 | 3.50 | 3.18 | 75% | B |
| 152010006 | Rulo Sembiring | 3.33 | 3.65 | 3.68 | 90% | B |
| 152010007 | Rani Perida Sitompul | 3.12 | 3.00 | 3.33 | 70% | B |
| 152010008 | Idaman Sitepu | 3.46 | 3.68 | 3.80 | 100% | A |
| 152010009 | Rika Melisa Br Sembiring | 3.55 | 3.72 | 3.90 | 100% | A |
| 152010010 | Bahagia Sitepu | 3.68 | 3.80 | 3.84 | 100% | A |
| 152010011 | Riko Johans Lubis | 2.90 | 3.12 | 3.20 | 70% | B |
| 152010012 | Lidia Megawati Br Kembaren | 3.68 | 3.72 | 3.69 | 95% | A |
| 152010013 | Pebrina Megsari Br Kembaren | 3.55 | 3.90 | 3.92 | 100% | A |
| 152010014 | Heliyani Welbri Br Sembiring | 3.72 | 3.55 | 3.60 | 100% | A |
| 152010015 | Keke Putri Br Pinem | 3.19 | 3.14 | 3.33 | 80% | B |
| 152010019 | Bermanta Sry Romario Karosekali | 3.24 | 3.12 | 3.00 | 80% | B |
| 152010020 | Gladis Lavenia Sembiring | 3.80 | 3.82 | 3.68 | 100% | A |
| 162010014 | Christian Alvinus | 3.00 | 3.12 | 3.55 | 90% | B |

### Table 3. Classification Accuracy between SSVM and Decision Tree

|                | SSVM         | Decision Tree |
|----------------|--------------|---------------|
| Best Parameter | 1.80, 3.37e-005 | 216.35, 0.16 |
| Training Accuracy (%) | 77.86 | 88.9 |
| Testing Accuracy (%)  | 76.82 | 86.9 |
| CPU Speed (Sec)      | 402.893 | 900.378 |

4. Conclusion

In conclusion, the results can be seen as follows:

a. Decision Tree performance is better than SSVM
b. Decision Tree gets very good results that promise to help find the best students to get scholarships
c. The accuracy of the Decision Tree classification in this study is better than SSVM

References

[1] Dai Qin-yun., Zang Chun-Ping., Wu Hao. 2016. Research of Decision tree Classification Algorithm in Data Mining. Dept. of Electric and Electronic Engineering, Shijiazhuang Vocational and Technology Institute. China.

[2] S. Sudirman, A. P. Windarto, and A. Wanto, “Data Mining Tools | RapidMiner: K-Means Method on Clustering of Rice Crops by Province as Efforts to Stabilize Food Crops In Indonesia,” IOP Conference Series: Materials Science and Engineering, vol. 420, no. 12089, pp. 1–8, 2018.

[3] A. Wanto, M. Zarlis, Sawaluddin, and D. Hartama, “Analysis of Artificial Neural Network Backpropagation Using Conjugate Gradient Fletcher Reeves in the Predicting Process,” Journal of Physics: Conference Series, vol. 930, no. 1, pp. 1–7, 2017.

[4] Hua CHEN., Jing GUO., Wei XIONG., Shenglian GUO., XU Chong-Yu. 2009. Downscaling GCMs Using the Smooth Support Vector Machine Method to Predict Daily Precipitation in the Hanjiang Basin. State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University. Wuhan.

[5] Lee, Y. J., & Mangasarian, O. L. 2001. RSVM: Reduced Support Vector Machine. In Proceedings of the First SIAM International Conference on Data Mining.

[6] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. P. Windarto, “Application of Classification Method C4.5 on Selection of Exemplary Teachers,” in IOP Conference Series,
[7] Lee, Y. J., & Mangasarian, O. L. 2000. SSVM: A Smooth Support Vector Machine for Classification. Computer Science Department, University of Wisconsin, Madison, Wisconsin.

[8] Patel, B. N., Prajapati G. Sathish., Lakhtaria I. Kamaljit. 2012. Efficient Classification of Data Using Decision Tree. Bonfring International Journal of Data Mining, Vol. 2, No.1.

[9] Purnami Santi Wulan., Embong Abdullah., Zain Jasni Mohd., and Rahayu S.P. 2009. A New Smooth Support Vector Machine and Its Application in Diabetes Disease Diagnosis. University Malaysia Pahang, Kuantan Pahang, Malaysia.

[10] B. Febriadi, Z. Zamzami, Y. Yunefri, and A. Wanto, “Bipolar function in backpropagation algorithm in predicting Indonesia’s coal exports by major destination countries,” IOP Conference Series: Materials Science and Engineering, vol. 420, no. 12089, pp. 1–9, 2018

[11] Purnami Santi Wulan., Andari Shofi., Pertiwi Yuniati Dian. 2015. High-Dimensional Data Classification Based on Smooth Support Vector Machine. Institut Teknologi Sepuluh Nopember, Surabaya.

[12] Manasi M. Phadatare, Sushma S. Nandgaonkar. 2014. Uncertain Data Mining using Decision Tree and Bagging Technique. Department of Computer Engineering, India.

[13] Seema., Rathi Monika., Mamta. 2012. Decision Tree: Data Mining Techniques. Department of Computer Science Engineering, India.

[14] Turban Efraim., E Jay., Aronson., Liang Ting-Peng. 2005. Decision Support System and Intelligent System. Andi Offset

[15] Vapnik, V. 1998. Statistical Learning Theory. Wiley New York

[16] Wu,Q., & Wang, W. 2013. Piecewise-Smooth Support Vector Machine for Classification. Mathematical problem in Engineering.

[17] Yuan Yubo., Fan Weigo., Pu Dongmei. 2007. Spline Function Smooth Support Vector Machine For Classification. University of Electronics Science and Technology of China Chengdu, China.