MODEL OPTIMISASI HYBRID ENSEMBLES DALAM MENYELESAIKAN PERMASALAHAN CLASS IMBALANCE

Hartono
Program Studi Teknik Informatika, STMIK IBBI
Jalan Sei Deli No. 18 Medan, Medan-Indonesia
E-mail : hartonoibbi@gmail.com

Abstrak

Objek penelitian yang akan diteliti pada penelitian ini adalah permasalahan class imbalance yang merupakan salah satu permasalahan utama di dalam klasifikasi data. Salah satu metode yang dapat dipakai untuk menyelesaikan permasalahan class imbalance adalah Hybrid Ensembles. Namun, terdapat 2 (dua) isu utama yang perlu dibahas yaitu masalah jumlah classifier dan diversity data. Peneliti di dalam penelitian akan melakukan proses optimisasi terhadap hybrid ensembles. Peneliti akan mengemukakan model yang menambahkan ensemble learning berbasis sample (sample based). Di dalam model ini peneliti akan melakukan penggabungan penggunaan bagging dengan metode DCS (Differenct Contribution Sampling) yang terbukti dapat mengurangi jumlah classifier dan meningkatkan diversity data dibandingkan dengan menggunakan AdaBoost. Model ini juga akan mengunakan Random Balance Ensemble Method yang menggabungkan random undersampling dengan SMOTEBoost yang dapat digunakan tahapan prepocessing yang ditujukan untuk mengurangi ukuran data training dan meningkatkan diversity dari dataset.

Kata Kunci : *Hybrid Ensembles, Ensemble Learning, Different Contribution Sampling, AdaBoost, SMOTEBoost*

Abstract

The object of research that will be examined in this study is the class imbalance problem which is one of the main problems in data

classification. One method that can be used to solve the problems of imbalance is a Hybrid Ensembles class. However, there are 2 (two) main issues that need to be discussed is the problem number classifier and diversity data. Researchers in the study will make the process of optimization of the hybrid ensembles. Researchers will propose a model that adds ensemble learning based sample (sample based). In this model the researchers will merge bagging method using DCS (differenct Contribution Sampling), shown to reduce the number of classifier and increase the diversity of data were compared using AdaBoost. This model will also use the Random Balance Method Ensemble which combines random undersampling with SMOTEBoost usable prepocessing stages aimed at reducing the size of the training data and increase the diversity of the dataset.

Keywords : ***Hybrid Ensembles, Ensemble Learning, Different Contribution Sampling, AdaBoost, SMOTEBoost***

1. PENDAHULUAN

Di dalam klasifikasi, *data-set* dikatakan sebagai tidak seimbang (*imbalanced*) ketika terdapat suatu *class* dengan jumlah data yang yang lebih kecil dibandingkan dengan *class* yang lain (Chawla *et al.*, 2004). *Class imbalance* merupakan suatu masalah karena suatu *machine learning* akan menghasilkan suatu akurasi prediksi klasifikasi yang baik terhadap kelas *data training* dengan jumlah anggota yang banyak, sedangkan kelas dengan jumlah anggota sedikit memiliki akurasi yang jelek (Galar *et al.*, 2012).

Galar *et al.* (2012) telah mengemukakan taksonomi yang terdiri-dari 4 (empat) kelompok pendekatan di dalam penggunaan metode *ensemble learning* yaitu: *cost-sensitive boosting*, *boosting-based ensembles*, *bagging-based ensembles*, dan *hybrid ensembles*. *Hybrid Ensembles*, merupakan algoritma yang menggabungkan *bagging* maupun *boosting* dan juga teknik *prepocessing*. *Hybrid Ensembles* menggunakan *bagging* sebagai metode *ensemble learning* dan di dalam tiap tahapan *bagging* terdapat metode *boosting*.

Galar *et al.* (2012) menggunakan metode *UnderBagging* dan *AdaBoost* di dalam taksonomi *hybrid ensembles* yang mereka kemukakan. Penelitian yang dilakukan oleh Galar *et al.* (2012) menunjukkan bahwa penggunaan metode *UnderBagging* dan *AdaBoost* memerlukan penggunaan jumlah *classifier* yang cukup besar. Metode *Preprocessing* yang digunakan adalah metode SMOTEBoost.

Seiffert *et al*. (2010) mengemukakan teori *RandomUndersampling Boost* (RUSBoost) yang merupakan merupakan gabungan dari penggunaan *RandomUndersampling* dan metode SMOTEBoost. Penggunaan metode RUSBoost dapat memberikan *performance* yang lebih baik dari *AdaBoost* dengan jumlah *classifier* yang jauh lebih sedikit.

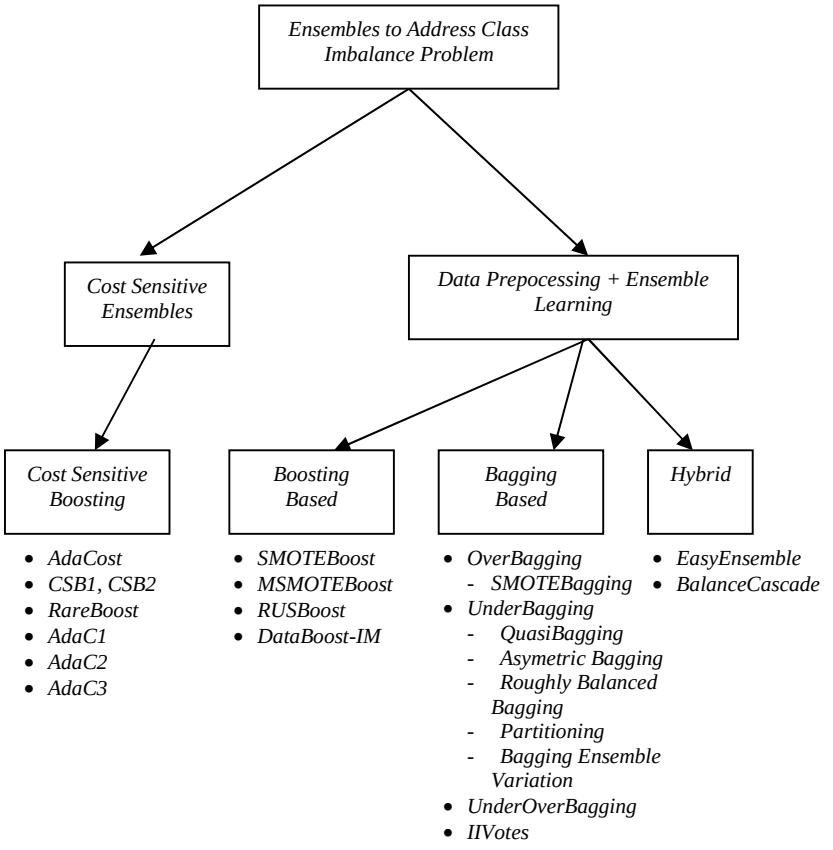
Taksonomi yang dikemukakan oleh Galar *et al*. (2012) telah dipakai secara luas oleh beberapa peneliti. Galar *et al.* (2013) berusaha menyempurnakan taksonomi, di mana dihadirkan pendekatan baru di dalam penggunaan metode *Boosting* dengan menggunakan pendekatan EUSBoost yang menggunakan metode RUSBoost, tetapi penelitian ini mengorbankan masalah keanekaragaman (*diversity*) data. Fernandez *et al.* (2013) menggunakan metode *hybrid ensembles* untuk menyelesaikan permasalahan *multi class* dengan menggabungkan penggunaannya dengan metode SVM untuk memecah permasalahan menjadi permsalahan *two class*, tetapi hasil penelitian masih menghasilkan akurasi yang rendah. Pengyi *et al*. (2014) di dalam penelitiannya mengenai pemanfaatan *ensemble learning* menggunakan metode *Sumple Subset Optimization* di dalam pengambilan *sample dataset* menunjukkan bahwa penggunaan *hybrid ensemble* perlu memperhatikan masalahan *diversity data*. Penelitian yang dilakukan oleh Krawczyk (2015) memodifikasi metode *hybrid ensemble* dan memasukkan proses *feature selection* pada tahap *bagging* dengan menggunakan *genetic algorithm*, tetapi metode yang mereka kemukakan gagal untuk menjawab masalah akurasi data dan hasil penelitian mereka menunjukkan bahwa pengukuran *diversity* perlu mendapat perhatian.

Jian *et al.* (2016) mengemukakan Metode *ensemble learning* yang baru yang dinamakan sebagai Different Contribution Sampling (DCS), yang dapat dikatakan sebagai metode *ensemble* yang berbasis *sampling* (*sampling-based*) dan *Boosting*. Di dalam metode DCS, *Biased Support Vector Machine* (B-SVM) digunakan untuk membangkitkan *Non-Support Vector* (NSV) dan *Support Vector* (SV). *Support Vector* (SV) yang dikombinasikan dengan SMOTE (SV-SMOTE) akan digunakan untuk meningkatkan jumlah anggota dari *minority class* dan *Non-Support Vector* yang dikombinasikan dengan RUSBoost (NSV-RUS) akan digunakan untuk menurunkan jumlah anggota dari *mayority class*. Penggunaa metode DCS berdasarkan hasil penelitian mereka dapat meningkatkan keanekaragaman (*diversity*) data.

Jose *et al*. (2015) mengemukakan metode *Random Balance Ensemble Method* yang menggabungkan *random undersampling* dengan SMOTEBoost yang dapat digunakan tahapan *prepocessing* yang ditujukan untuk mengurangi ukuran *data training* dan meningkatkan *diversity* dari *dataset*.

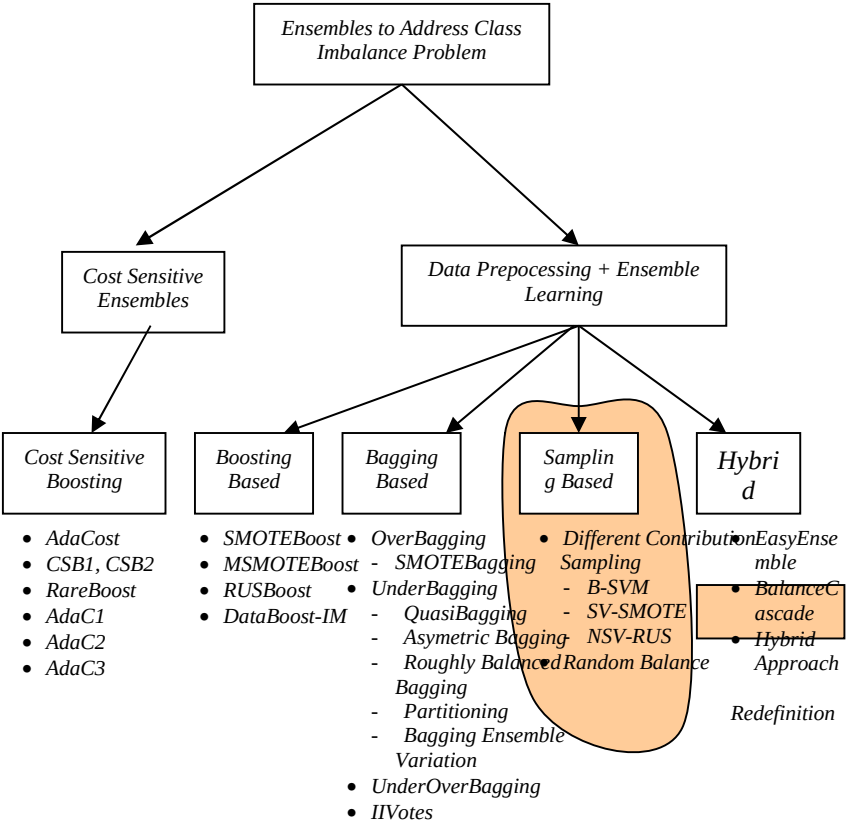
Peneliti di dalam penelitian ini mengemukakan model model taksonomi baru yang memodifikasi model taksonomi yang dikemukakan oleh Galar *et al.* (2012). Di dalam taksonomi yang dihadirkan oleh peneliti dihasilkan pendekatan baru yaitu *sampling-based*. Peneliti juga menghadirkan *model* Optimisasi *Hybrid Ensembles* yang menggabungkan pemakaian *bagging* dengan metode DCS. Di dalam metode DCS sendiri sebenarnya merupakan *ensemble* dari metode *sampling* dengan metode *Boosting*. Terdapat 2 (dua) isu utama yang akan dijawab yaitu mengenai masalah jumlah *classifier* dan *divesity* data. Metode *prepocessing* yang akan digunakan adalah metode *Random Balance Ensemble Method*. Penggunaan metode DCS yang digabungkan dengan *UnderBagging* diharapkan dapat mengurangi jumlah *classifier* dan juga meningkatkan *diversity* data.

Adapun Model Taksonomi Galar yang selengkapnya dapat dilihat pada Gambar 1.



Gambar 1. Taksonomi Galar untuk *Ensemble Learning*

Adapun model taksonomi yang diusulkan oleh peneliti dapat dilihat pada Gambar 2.2.

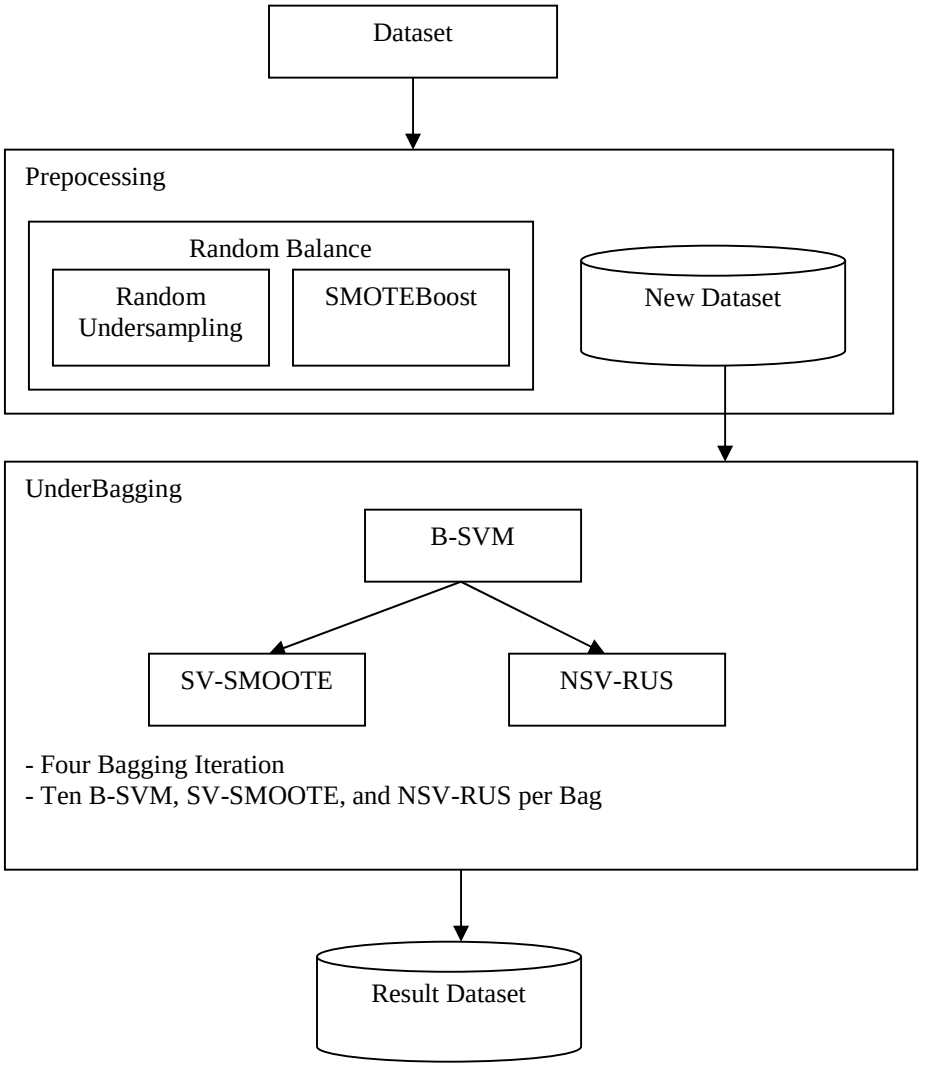


Gambar 2. Model Taksonomi yang Diusulkan oleh Peneliti

2. METODE PENELITIAN

Data yang digunakan merupakan 44 *binary data-sets* yang bersumber dari KEEL *data-set repository*. *Multi class data-sets* akan diubah menjadi permasalahan *two-class* dengan cara mengelompokkan *class* menjadi *positive class* dan *negative class*.

Adapun prosedur kerja yang dilakukan oleh peneliti dari penelitian ini dapat dilihat secara keseluruhan pada Gambar 3.



Gambar 3. Tahapan Metode Penelitian

3. HASIL DAN PEMBAHASAN

3.1. *Prepocessing* dengan Menggunakan *Random Balance Ensemble Method*

Di dalam taksonomi Galar (2012), metode *hybrid ensembles* menggunakan SMOTEBoost sebagai metode *prepocessing*. Jose *et al*. (2015) mengemukakan metode *Random Balance Ensemble Method* yang menggabungkan *random undersampling* dengan SMOTEBoost yang dapat digunakan tahapan *prepocessing* yang ditujukan untuk mengurangi ukuran *data training* dan meningkatkan *diversity* dari *dataset*.

Adapun *pseudocode* dari metode *Random Ensemble Method* adalah sebagai berikut.

Require: Set S of examples (x_1, y_1), ... , (x_m, y_m) where $x_1 \varepsilon$ X and $y_i \varepsilon$ Y = {-1, +1} (+1: positive or minority class, -1: negative or majority class), neighbours used in SMOTE, *k*

Ensure: New set *S′* of examples with Random Balance

1: *totalSize* ← |S|
2: $S_N \leftarrow \{(x_i, y_i) \varepsilon S \mid y_i = -1\}$
3: $S_P \leftarrow \{(x_i, y_i) \varepsilon S \mid y_i = +1\}$
4: *majoritySize* ← $|S_N|$
5: *minoritySize* ← $|S_P|$
6: *newMajoritySize* ← Random integer between 2 and *totalSize*-2
7: *newMinoritySize* ← *totalSize* – *newMajoritySize*
8: **if** *newMajoritySize* < *majoritySize* **then**
9: $S' \leftarrow S_P$
10: Take a random sample of size *newMajoritySize* from S_N, add the sample to *S′*
11: Create *newMinoritySize* – *minoritySize* artificial
12: **else**
13: $S' \leftarrow S_N$
14: Take a random sample of size *newMinoritySize* from S_P, add the sample to *S′*
15: create *newMajoritySize* – *majoritySize* artificial
16: **end if**

17: **return** S'

3.2. *Different Contribution Sampling* (DCS)

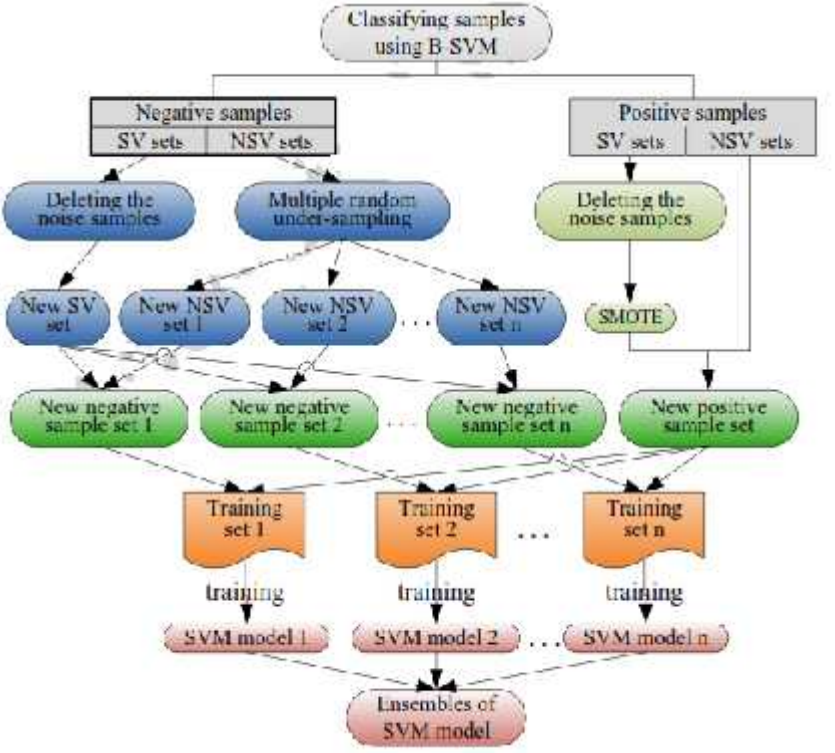
Jian *et al.* (2016) mengemukakan Metode *ensemble learning* yang baru yang dinamakan sebagai Different Contribution Sampling (DCS), yang dapat dikatakan sebagai metode *ensemble* yang berbasis *sampling* (*sampling-based*) dan *Boosting*. Di dalam metode DCS, *Biased Support Vector Machine* (B-SVM) digunakan untuk membangkitkan *Non-Support Vector* (NSV) dan *Support Vector* (SV). *Support Vector* (SV) yang dikombinasikan dengan SMOTE (SV-SMOTE) akan digunakan untuk meningkatkan jumlah anggota dari *minority class* dan *Non-Support Vector* yang dikombinasikan dengan RUSBoost (NSV-RUS) akan digunakan untuk menurunkan jumlah anggota dari *mayority class*. Penggunaa metode DCS berdasarkan hasil penelitian mereka dapat meningkatkan keanekaragaman (*diversity*) data.

Adapun langkah-langkah di dalam mengimplementasikan metode DCS adalah sebagai berikut.

1. *Support Vector* dan *non Support Vector* dari *minority* dan *majority* adalahh diidentifikasikan melalui metode *Biased Support Vector Machine* (B-SVM).
2. *Minority class* yang dihasilkan melalui metode SMOTE diperoleh melalui pemangkasan *Support Vector* pada *Minority*; *Minority dataset* terdiri-dari *Support Vector* yang baru dan *Non Support Vector* yang asli pada *minority*.
3. *Multiple Random Under Sampling (RUS) Boost* digunakan untuk mengeliminasi *Non Support Vector* (NSV) dari *Majority* untuk memperoleh *Multiple Sets* dari NSV baru, selagi *noise* dari *Support Vector* dari *majority* dihilangkan; *Multiple Majority Datasets* terdiri-dari *Support Vector* yang telah dihilagkan *noise* dan *multiple sets* dari NSV baru di dalam *Majority*.
4. *Multiple Balanced Training Sets* dapat diperoleh melalui kombinasi dari *minority dataset* yang baru dengan *multiple new*

majority datasets.

Adapun proses kerja dari metode DCS dapat dilihat pada Gambar 4.



Gambar 4. Model Kerja dari Metode DCS

3.3. Hasil Penelitian

Berdasarkan hasil penelitian yang dilakukan oleh peneliti dapat diperoleh bahwa Metode *hybrid ensembles* sebenarnya sudah cukup baik. Namun, terdapat 2 (dua) isu utama yang perlu

dibahas yaitu masalah jumlah *classifier* dan *diversity data*. Hal itu telah disadari oleh Galar *et al.* Dalam penelitiannya pada tahun 2013, yang menggunakan metode EUSBoost dengan menerapkan RUSBoost yag memang dapat mengurangi jumlah *classifier*, tetapi masalah *diversity data* belum terpecahkan. Berdasarkan hasil penelitian yang dilakukan oleh sejumlah peneliti, maka peneliti di dalam penelitian ini akan mengemukakan model optimisasi terhadap *Hybrid Ensembles* dan dapat menjawab isu yang ada. Pemakaian Metode DCS yang digabungkan dengan *UnderBagging* dapat mengurangi jumlah *classifier* dan juga menjaga *diversity* data, demikian juga *prepocessing* dengan menggunakan *Random Balance Ensemble Method* juga dapat mengurangi ukuran *data training* dan meningkatkan *diversity* data. Di dalam metode DCS, *Biased Support Vector Machine* (B-SVM) digunakan untuk membangkitkan *Non-Support Vector* (NSV) dan *Support Vector* (SV). *Support Vector* (SV) yang dikombinasikan dengan SMOTE (SV-SMOTE) akan digunakan untuk meningkatkan jumlah anggota dari *minority class* dan *Non-Support Vector* yang dikombinasikan dengan RUSBoost (NSV-RUS) akan digunakan untuk menurunkan jumlah anggota dari *mayority class*. Penggunaa metode DCS berdasarkan hasil penelitian dapat meningkatkan keanekaragaman (*diversity*) data.

3.4. Pembahasan

Penelitian ini hanya digunakan pada *data-set repository* yang merupakan *data set* dengan permasalahan *binary class*. Hasil penelitian di masa mendatang hendaknya dapat dilakukan pada *Multi Class Data-Sets*. *Multi class data-sets* akan diubah menjadi permasalahan *two-class* dengan cara mengelompokkan *class* menjadi *positive class* dan *negative class*.

4. KESIMPULAN

Adapun kesimpulan yang dapat diperoleh dari hasil penelitian adalah sebagai berikut.

1. Secara umum terdapat 2 (dua) permasalahan utama dari *Hybrid Ensembles* di dalam menyelesaikan permasalahan *class imbalance* yaitu masalah jumlah *classifier* dan *diversity data*.
2. *Prepocessing* dengan menggunakan *Random Balance Ensemble Method* yang menggabungkan *Random Under Sampling* dan *SMOOTEBoost* dapat mengurangi ukuran *data training* dan meningkatkan *diversity* dari *dataset*.
3. Metode *Different Contribution Sampling* (DCS) yang merupakan pendekatan *sampling based* dapat menyempurnakan *Hybrid Ensembles*. Di dalam metode DCS, *Biased Support Vector Machine* (B-SVM) digunakan untuk membangkitkan *Non-Support Vector* (NSV) dan *Support Vector* (SV). *Support Vector* (SV) yang dikombinasikan dengan SMOTE (SV-SMOTE) akan digunakan untuk meningkatkan jumlah anggota dari *minority class* dan *Non-Support Vector* yang dikombinasikan dengan RUSBoost (NSV-RUS) akan digunakan untuk menurunkan jumlah anggota dari *mayority class*. Penggunaan metode DCS berdasarkan hasil penelitian dapat meningkatkan keanekaragaman (*diversity*) data.

5. SARAN

Saran untuk penelitian di masa mendatang adalah penelitian di masa mendatang hendaknya dapat dilakukan pada *Multi Class Data-Sets*.

DAFTAR PUSTAKA

Chawla, N.V., N. Japkowicz, and A. Kolcz. 2004. Special Issue Learning Imbalanced Datasets. *SGIKDD Explor. Newsl* **6**(1)

Fernandez, A., V. Lopez, M. Galar, M.J.D. Jesus, F. Herrera. 2013. Analysing the Classification of Imbalanced Data-Sets with Multiple Classes: Binarization Techniques and Ad-hoc Approaches. *Knowledge-Based Systems* **42**: 97-110

Galar, M., A. Fernandez, E. Barrenechea, and H. Bustince. 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging, Boosting, and Hybrid-Based Approachs. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews***42**(4): 1-21

Galar, M., A. Fernandez, E. Barrenechea, and F. Herrera. 2013. EUSBoost: Enhancing Ensembles for Highly Imbalanced Data-Sets by Evolutionary Undersampling. *Pattern Recognition* **46**: 3460-3471

Jian, C., J. Gao, and Y. Ao. 2016. A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble. *Neurocomputing*

Jose, F.D.P., J.J. Rodriguez, C.G. Osorio, and L.I. Kuncheva. 2015. Random Balance: Ensembles of Variable Priors Classifiers for Imbalanced Data. *Knowledge-Based Systems***85**(2015): 96-111

Krawczyk, B., G. Schaefer, M. Wozniak. 2015. A Hybrid Cost-Senstivie Ensemble for Imbalanced Breast Thermogram Classification. *Artificial Intelligence in Medicine***65** (2015): 219-227

Pengyi, Y. *et al.* 2014. Sample Subset Optimization Techniques for Imbalanced and Ensemble Learning Problems in Bioinformatics Applications. IEEE Transaction on Cybernetics**44** (3): 445-455

Seiffert, C., T. Khoshgoftaar, J.V. Hulse, and A. Napolitano. 2010. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Trans. Syst. Man. Cybern. A. Syst* **40**(1): 185-197