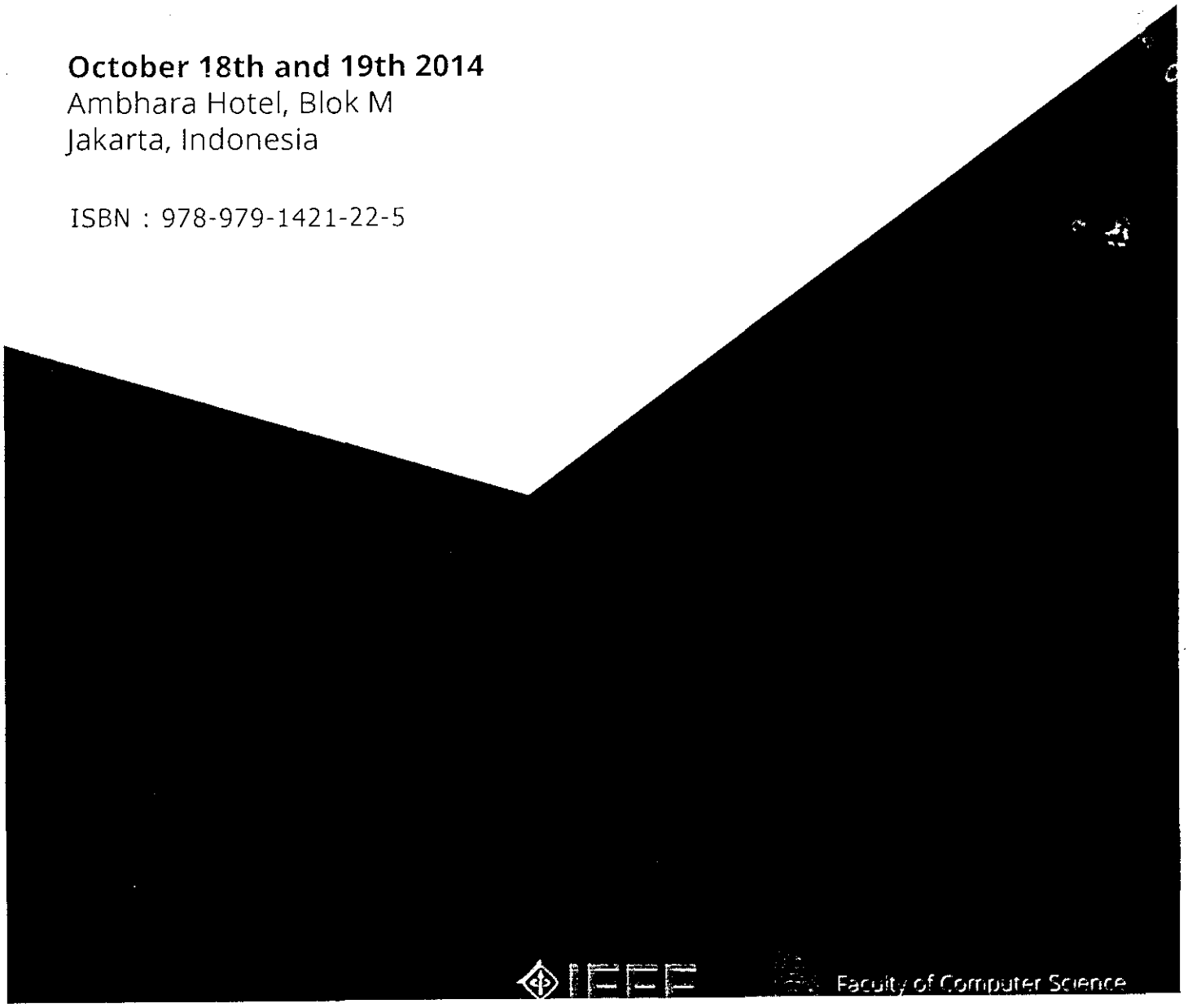
ICACSIS 2014

2014 International Conference on Advanced Computer Science and Information Systems

October 18th and 19th 2014
Ambhara Hotel, Blok M
Jakarta, Indonesia

ISBN : 978-979-1421-22-5



IEEE INDONESIA SECTION

ICACSIS 2014

2014 International Conference on

Advanced Computer Science and Information Systems

(Proceedings)

ISBN : 978-979-1421-225

Welcome Message from General Chairs

On behalf of the Organizing Committee of this International Conference on Advanced Computer Science and Information Systems 2014 (ICACSIS 2014), we would like to extend our warm welcome to all of the presenter and participants, and in particular, we would like to express our sincere gratitude to our plenary and invited speakers.

This international conference is organized by the Faculty of Computer Science, Universitas Indonesia, and is intended to be the first step towards a top class conference on Computer Science and Information Systems. We believe that this international conference will give opportunities for sharing and exchanging original research ideas and opinions, gaining inspiration for future research, and broadening knowledge about various fields in advanced computer science and information systems, amongst members of Indonesian research communities, together with researchers from Germany, Singapore, Thailand, France, Algeria, Japan, Malaysia, Philippines, United Kingdom, Sweden, United States and other countries.

This conference focuses on the development of computer science and information systems. Along with 4 plenary and 2 invited speeches, the proceedings of this conference contains 71 papers which have been selected from a total of 132 papers from twelve different countries. These selected papers will be presented during the conference.

We also want to express our sincere appreciation to the members of the Program Committee for their critical review of the submitted papers, as well as the Organizing Committee for the time and energy they have devoted to editing the proceedings and arranging the logistics of holding this conference. We would also like to give appreciation to the authors who have submitted their excellent works to this conference. Last but not least, we would like to extend our gratitude to the Ministry of Education of the Republic of Indonesia, the Rector of Universitas Indonesia, Universitas Tarumanagara, Bogor Agricultural Institute, and the Dean of the Faculty of Computer Science for their continued support towards the ICACSIS 2014 conference.

Sincerely yours,
General Chairs

ISBN : 978-979-1421-225

Welcome Message from The Dean of Faculty of Computer Science, Universitas Indonesia

On behalf of all the academic staff and students of the Faculty of Computer Science, Universitas Indonesia, I would like to extend our warmest welcome to all the participants to the Ambhara Hotel, Jakarta on the occasion of the 2014 International Conference on Advanced Computer Science and Information Systems (ICACSIS).

Just like the previous five events in this series (ICACSIS 2009, 2010, 2011, 2012, and 2013), I am confident that ICASIS 2014 will play an important role in encouraging activities in research and development of computer science and information technology in Indonesia, and give an excellent opportunity to forge collaborations between research institutions both within the country and with international partners. The broad scope of this event, which includes both theoretical aspects of computer science and practical, applied experience of developing information systems, provides a unique meeting ground for researchers spanning the whole spectrum of our discipline. I hope that over the next two days, some fruitful collaborations can be established.

I also hope that the special attention devoted this year to the field of pervasive computing, including the very exciting area of wireless sensor networks, will ignite the development of applications in this area to address the various needs of Indonesia's development.

I would like to express my sincere gratitude to the distinguished invited speakers for their presence and contributions to the conference. I also thank all the program committee members for their efforts in ensuring a rigorous review process to select high quality papers.

Finally, I sincerely hope that all the participants will benefit from the technical contents of this conference, and wish you a very successful conference and an enjoyable stay in Jakarta.

Sincerely,
Mirna Adriani, Dra, Ph.D.
Dean of the Faculty of Computer Science
Universitas Indonesia

ISBN : 978-979-1421-225

COMMITTEES

Honorary Chairs:

- A. Jain, IEEE Fellow, Michigan State University, US
- T. Fukuda, IEEE Fellow, Nagoya University, JP
- M. Adriani, Universitas Indonesia, ID

General Chairs:

- E. K. Budiarjo, Universitas Indonesia, ID
- D. I. Sensuse, Universitas Indonesia, ID
- Z. A. Hasibuan, Universitas Indonesia, ID

Program Chairs:

- H. B. Santoso, Universitas Indonesia, ID
- W. Jatmiko, Universitas Indonesia, ID
- A. Buono, Institut Pertanian Bogor, ID
- D. E. Herwindiati, Universitas Tarumanegara, ID

Section Chairs:

- K. Wastuwibowo, IEEE Indonesia Section, ID

Publication Chairs:

- A. Wibisono, Universitas Indonesia, ID

Program Committees:

- A. Azurat, Universitas Indonesia, ID
- A. Fanar, Lembaga Ilmu Pengetahuan Indonesia, ID
- A. Kistijantoro, Institut Teknologi Bandung, ID
- A. Purwarianti, Institut Teknologi Bandung, ID
- A. Nugroho, PTIK BPPT, ID
- A. Srivihok, Kasetsart University, TH
- A. Arifin Institut Teknologi Sepuluh Nopember, ID
- A. M. Arymurthy, Universitas Indonesia, ID
- A. N. Hidayanto, Universitas Indonesia, ID
- B. Wijaya, Universitas Indonesia, ID
- B. Yuwono, Universitas Indonesia, ID
- B. Hardian, Universitas Indonesia, ID
- B. Purwandari, Universitas Indonesia, ID
- B. A. Nazief, Universitas Indonesia, ID
- B. H. Widjaja, Universitas Indonesia, ID
- Denny, Universitas Indonesia, ID
- D. Jana, Computer Society of India, IN
- E. Gaura, Coventry University, UK
- E. Seo, Sungkyunkwan University, KR
- F. Gaol, IEEE Indonesia Section, ID

ISBN : 978-979-1421-225

- M. I. Suryani, Universitas Indonesia, ID
- M. Soleh, Universitas Indonesia, ID
- M. Pravitasari, Universitas Indonesia, ID
- N. Rahmah, Universitas Indonesia, ID
- Putut, Universitas Indonesia, ID
- R. Musfikar, Universitas Indonesia, ID
- R. Latifah, Universitas Indonesia, ID
- R. P. Rangkuti, Universitas Indonesia, ID
- V. Ayumi, Universitas Indonesia, ID

ISBN : 978-979-1421-225

CONFERENCE INFORMATION

Dates	October 18th (Saturday) – October 19th (Sunday) 2014
Organizer	Faculty of Computer Science, Universitas Indonesia
	Department of Computer Science, Institut Pertanian Bogor
	Faculty of Information Technology, Universitas Tarumanegara
Venue	Ambhara Hotel
	Jalan Iskandarsyah Raya No. 1, Jakarta Selatan, DKI Jakarta, 12160, Indonesia
	Phone : +62-21-2700 888
	Fax : +62-21-2700 215
	Website : http://www.ambharahotel.com/
Official Language	English
Secretariat	Faculty of Computer Science, Universitas Indonesia
	Kampus UI Depok
	Depok, 16424
	Indonesia
	T: +62 21786 3419 ext. 3225
	F: +62 21 786 3415
	E: icacsis@cs.ui.ac.id
	W: http://www.cs.ui.ac.id
Conference Website	http://icacsis.cs.ui.ac.id

ISBN : 978-979-1421-225

ADVANCED PROGRAM *ICACSIS 2014*

Sunday, October 19th, 2014-CONFERENCE			
Time	**Event**	**Event Details**	**Rooms**
08.00-09.00	Registration		Dirgantara Room, 2nd Floor
09.00-10.00	Plenary Speech III	Drs. Harry Waluyo, M.Hum from Directorate General of Media, Design, Science & Technology Based Creative Economy	
10.00-10.15	Coffee Break		
10.15-11.30	Plenary Speech IV	Prof. Masatoshi Ishikawa from University of Tokyo, JP	
11.30-12.30	Lunch		
12.30-14.00	Parallel Session IV : Four Parallel Sessions	**See Technical (Parallel Session IV Schedule)**	Elang, Kasuari, Merak, Cendrawasih Room, Lobby Level
14.00-15.30	Parallel Session V : Four Parallel Sessions	**See Technical (Parallel Session V Schedule)**	Elang, Kasuari, Merak, Cendrawasih Room, Lobby Level
15.30-16.00	Coffee Break		
16.00-16.30	Closing Ceremony	Awards Announcement and Photo Session	Dirgantara Room, 2nd Floor

ISBN : 978-979-1421-225

Table of Contents

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

Information Retrieval

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

ISBN : 978-979-1421-225

Identification of Single Nucleotide Polymorphism using Support Vector Machine on Imbalanced Data

Lailan Sahrina Hasibuan, Wisnu Ananta Kusuma, Willy Bayuardi Suwarno
Department of Computer Science, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University
Department of Agronomy and Horticulture, Faculty of Agriculture, Bogor Agricultural University
E-mail: lailan.sahrina.hsb@gmail.com, ananta@ipb.ac.id, bayuardi@gmail.com

Abstract*—The advance of DNA sequencing technology presents a significant bioinformatic challenges in a downstream analysis such as identification of single nucleotide polymorphism (SNP). SNP is the most abundant form of genetic marker and have been one of the most crucial researches in bioinformatics. SNP has been applied in wide area, but analysis of SNP in plants is very limited, as in cultivated soybean (Glycine max L.). This paper discusses the identification of SNP in cultivated soybean using Support Vector Machine (SVM). SVM is trained using positive and negative SNP. Previously, we performed a balancing positive and negative SNP with undersampling and oversampling to obtain training data. As a result, the model which is trained with balanced data has better performance than that with imbalanced data.

Keywords- identification; SNP; SVM; oversampling; undersampling.

I. INTRODUCTION

THE latest technological advancement in DNA sequencing, is the next generation of sequencing (NGS). NGS is also referred as high-throughput DNA sequencing (HTS). NGS or HTS has reduced the cost and increased the throughput of DNA sequencing because its ability to automate techniques in DNA sequencing [1]. However, NGS has shortcomings in the quality of the sequencing results, it produce shorter read lengths than previous technology, Sanger method [2]. It has a tremendous impact on how the reads have to be processed in a downstream analysis such as identification of Single Nucleotide Polymorphism (SNP).

SNP is the most abundant form of genetic marker in all populations of individuals [3]. SNP is defined as single base variations or short insertions or deletions in the nucleotide sequence from different individuals or between homologous sequences within an individual [4].

NGS technologies for SNP analysis has been applied in wide area such as medicine [5], animal genetics [6] and plant breeding [7]. However, SNP analysis in plants such as cultivated soybean (*Glycine max L.*) is limited [3].

This paper will focus on identification of SNP in cultivated soybean using Support Vector Machine (SVM). SVM is an effective method for general purpose supervised pattern recognition and has been applied successfully to many biological problems recently, such as gene selection [8], cancer classification [9] and detection of cardiac abnormality [10].

This paper is organized as follows. Section II will discuss about research related to identification of SNP that have been done before. Section III presents techniques at data level for pre-process, features used to train SVM, techniques to handling imbalance of data, undersampling and oversampling, training and testing with SVM. Results, discussion and conclusions are presented in section IV, V and VI.

II. RELATED WORK

Research about identification of SNP have been done both on DNA of human, animals and plants. In [4] they developed a machine learning to identify SNP on DNA of cultivated soybean. They used decission tree as classifier and feed forward neural network as learning method. The training data were 27.275 candidate SNP generated by sequencing 1973 STS (Sequence Tag Sites) from 6 diverse homozygous soybean cultivar. The accuracy reaches 97.3% with PPV of 84.8%.

Kong and Choo developed a method based on SVM (Support Vector Machine) to predict SNP. The research was conducted on human DNA. One-thousand positive SNP and 1000 negative SNP were randomly selected from JSNP, SNP database of the

Manuscript received July 28, 2014.

Lailan S. H and W. A. Kusuma are with the Department of Computer Science , Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Bogor 16680 (e-mail: lailan.sahrina.hsb@ gmail.com, ananta@ipb.ac.id).

W. B. Suwarno is with Department of Agronomy and Horticulture, Faculty of Agriculture, Bogor Agricultural University, Bogor 16680 (e-mail: bayuardi@gmail.com).

ISBN : 978-979-1421-225

Sampling outcome data, undersampling and oversampling, were divided into k subsets (k = 10). Of the k subsets, a single subset was retained as the testing data, and the remaining k − 1 subsets were used as training data. The cross-validation process is then repeated k times, with each of the k subsets used exactly once as the testing data. We did undersampling process two times with a value of m = 1 and m = 2.

We also conducted experiments on imbalanced data. We divided the data into k subsets (k = 10). Of the k subsets, a single subset was retained as the testing data, and the remaining k − 1 subsets were used as training data. The cross-validation process is then repeated k times, with each of the k subsets used exactly once as the testing data.

IV. RESULT

Identification of SNP that we did are divided into four categories: (1) identification of SNP using undersampling data with m = 1, (2) m = 2, (3) oversampling data, and (4) imbalance data. We used metrics based on confusion matrix as shown in Figure1 for evaluation. Confusion matrix has four categories: True positives (TP) are examples correctly labeled as positives. False positives (FP) refer to negative examples incorrectly labeled as positive. True negatives (TN) correspond to negatives correctly labeled as negative. Finally , False negatives (FN) refer to positive examples incorrectly labeled as negative [21].

We plotted the FPR (False Positive Rate) and TPR (True Positive Rate) of each categories (Figure 2). From this plotted we can see that category (1), (2), and (3) have better performance than category (4) because they have higher TPR and lower FPR. Generally, when TPR is high, then FPR is also high. This condition means that when a classifier has high ability in indentifying positive SNP as positive SNP then this classifier also has high fault in identifying negative SNP as positive SNP. The two criteria, TPR and FPR, are trade-off, we cannot use one of them to evaluate performance of a classifier. Hence we need a metric that combine the ability of classifier in indentifying positive SNP as positive SNP and the ability of

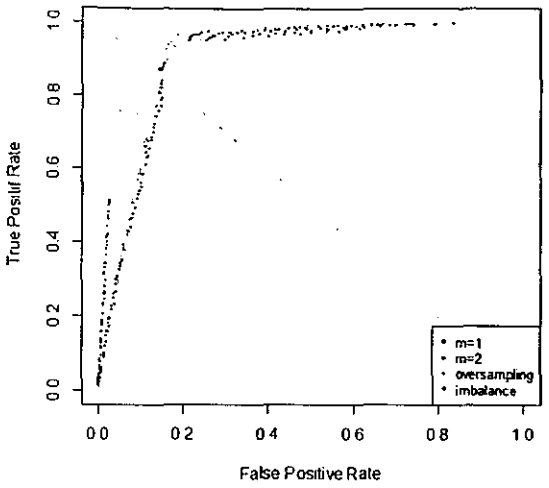


Fig 1. Plotted of false potive rate and true positive rate. Black dots shows the results of identification using training data undersampled with m = 1, blue dots is m = 2, green dots is oversampling, and red dots is imbalance data

classifier to not identify negative SNP as positive SNP. We used $F_{measure}$ to measure this metric, precision is refer to PPV and recal is refer to TPR.

$$F_{measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2)$$

We also measured G-Mean, the geometric mean of classifier performance on two classes separately.

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP} \quad (3)$$

$$G - Mean = \sqrt{sensitifity * specificity}$$

From each category, we choosed the best classifier based on $F_{measure}$. The best classifier of each category is presented in Table II. FPR and TPR plotted of best classifier, is presented in Figure 3.

	Prediction	
Target	T	F
T	TP	FN
F	FP	TN

(a) Confusion Matrix

True Positive Rate (TPR)	$\frac{TP}{(TP+FN)}$
False Positive Rate (FPR)	$\frac{FP}{(FP+TN)}$
Positive Predictive Value (PPV)	$\frac{TP}{(TP+FP)}$
Negative Predictive Value (NPV)	$\frac{TN}{(FN+TN)}$
Accuracy	$\frac{TP+TN}{(TP+FN+FP+TN)}$

(b) Definitions of Metrics

Fig 2. Common machine learning evaluation metrics

ISBN : 978-979-1421-225

Mean base quality (#9): mean quality of all bases at the variant position.

Alignment depth (#10): number of reads aligned at variant position.

Error probability (#11): probability of the number of the reads containing variant base was sampled from binomial distribution with given parameters.

Dinucleotide repeat (#12,13): number of dinucleotide repeat around the variant position (left and right direction).

Mismatch area (#14): mean number of variant base per each reads aligned at variant position.

Homopolymer length (#15,16): length of repeating bases around the variant position (left and right direction).

Nucleotide diversity (#17): deviation of reference base frequencies from given whole-genome average.

Total mismatch count (#18,19): number of mismatch on reads with reference base and reads with variant base.

Allele balance (#20): number of reads containing variant bases divided by alignment depth.

Mean of nearby base quality (#21): mean quality of all bases around variant position.

Distance to nearest variant (#22,23): distance of the variant to its neighboring variant (left and right flank size).

ACKNOWLEDGMENT

The authors would like to thank to Abrar Istiadi, a member of the Laboratory of Bioinformatics, Department of Computer Science, Bogor Agricultural University (IPB), for his very helpful assistance on the extraction of SNP features.

REFERENCES

[1] M. L. Metzker, "Sequencing technologies - the next generation," *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, Jan. 2010.

[2] Elaine R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends Genet*, vol. 24, no. 3, pp 133-141, Feb. 2008

[3] S. Kumar, T. W. Banks, S. Cloutier, "SNP discovery through Next-Generation Sequencing and its applications," *Int J Plant Genomics*, vol 2012, Nov. 2012

[4] L. K. Matukumalli, J. J. Grefenstette, D. L. Hyten, I.-Y. Choi, P. B. Cregan, and C. P. Van Tassell, "Application of machine learning in SNP discovery," *BMC Bioinformatics*, vol. 7, no. 4, Jan. 2006

[5] Nikolina Babić, "Clinical pharmacogenomics and concept of personalized medicine," *J Med Biochem*, vol. 31, no.4, pp 281-6.

[6] A. Vignal, D Milan, M. S. Cristobal, A. Eggen, "A review on SNP and other types of molecular markers and their use in animal genetics," *Genet. Sel. Evol*, vol. 34, no. 3, pp. 275-305, Jun 2002

[7] J. Mammadov, R. Aggarwal, R. Buyyarapu, and S. Kumpatla, "SNP markers and their impact on plant breeding," *Int. J. Plant Genomics*, vol. 2012, Jan. 2012.

[8] X. Zhou and D. P. Tuck, "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. " *BioinformaticsJ.* vol. 23, no. 9, pp. 1106–14, Sep. 2007.

[9] A. Bharathi and A.M. Natarajan, "Cancer classification using support vector machines and relevance vector machine based on analysis of variance features." *J. Computer Sci*, vol. 7, no. 9, pp. 1393-9, Sep. 2011.

[10] S. Ari, K. Hembram, G. Saha, "Detection of cardiac abnormality from PCG signal using LMS based least square SVM classifier." *Expert Syst Appl*, vol. 37, pp. 8019–26, Dec. 2010.

[11] W. Kong and K. W. Choo, "Predicting single nucleotide polymorphisms (SNP) from DNA sequence by support vector machine." *Front Biosci*, pp. 1610-4, Jan. 2007.

[12] B. D. O'Fallon, W. Wooderchak-Donahue, and D. K. Crockett, "A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data." *Bioinformatics*, vol. 29, no. 11, pp. 1361–6, Jun. 2013.

[13] H. Xiong, *et.al*, "A hybrid Fisher/SVM method for SNP discovery in Brassica oilseed rape." *J Food Agr Environ*, vol.8, no. 3&4, pp. 705-8, Okt. 2010.

[14] H.M. Lam, *et al.*, "Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection." *Nat. Genet.*, vol. 42, no. 12, pp. 1053–9, Dec. 2010.

[15] J. Schmutz, *et al.*, "Genome sequence of the palaeopolyploid soybean.," *Nature*, vol. 463, no. 7278, pp. 178–83, Jan. 2010.

[16] R. Li, *et al.*, "SOAP2: an improved ultrafast tool for short read alignment." *Bioinformatics*, vol. 25, no. 15, pp. 1966–7, Jun 2009.

[17] J. Van Oeveren and A. Janssen, "Mining SNPs from DNA Sequence Data: Computational Approaches to SNP Discovery and Analysis," in *Single Nucleotide Polymorphisms*, vol. 578, A. A. Komar, Ed. Totowa, NJ: Humana Press, 2009, pp. 73–91.

[18] S. J. Yen, Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions." *Expert Syst Appl*, vol. 36, pp. 5718-27, Apr. 2009.

[19] R. Batuwita, and V. Palade, "Efficient resampling methods for training support vector machines with imbalanced datasets." In *Neural Networks (IJCNN). The 2010 International Joint Conference on*, pp. 1-8. IEEE. 2010.

[20] D. Meyer, *et al.* e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. http://CRAN.R-project.org/package=e1071. 2014.

[21] Y. Tang, *et.al*. "SVMs modeling for highly imbalanced classification." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions* vol. 39, no. 1, pp. 281-288. Jan. 2009.

[22] R . Batuwita, and V. Palade, "Class imbalance learning methods for support vector machines" In: He, H., Ma, Y. (Eds.), Imbalanced Learning: Foundations, Algorithms, and Applications, in press, Wiley, 2013

ISBN : 978-979-1421-225