ClusterOSS: a new undersampling method for imbalanced learning

Brazilian Conference on Intelligent Systems, 3th; Encontro Nacional de Inteligência Artificial e Computacional, 11th, 2014, São Carlos.
http://www.producao.usp.br/handle/BDPI/48665

Downloaded from: Biblioteca Digital da Produção Intelectual - BDPI, Universidade de São Paulo
ENIAC 2014
Encontro Nacional de Inteligência Artificial e Computacional

19 - 23 Outubro 2014
São Carlos, São Paulo, Brasil
Preface

The joint conference Brazilian Conference on Intelligent Systems (BRACIS) – Encontro Nacional de Inteligência Artificial e Computacional (ENIAC) is the combination of the most important scientific events in the country historically related to Artificial Intelligence (AI) and Computational Intelligence (CI): the Brazilian Symposium on Artificial Intelligence (SBIA, 21 editions), the Brazilian Symposium on Neural Networks (SBRN, 12 editions) and the Encontro Nacional de Inteligência Artificial (ENIA, 10 editions).

The main goal of BRACIS is the promotion of international level research by exchanging scientific ideas among researchers, practitioners, scientists and engineers, as well as promoting original theories and novel applications dealing with the use and analysis of Artificial and Computational intelligence techniques in various related fields. ENIAC is a national forum for researchers, practitioners, educators and students to present and discuss innovations, trends, experiences, developments and work in progress in the fields of Artificial and Computational Intelligence. BRACIS ENIAC 2014 shares the same goals of its previous editions and adds the aim to stimulate the development of high-level research that crosses boundaries between Artificial and Computational Intelligence. BRACIS-ENIAC 2014 will be held in the intellectually vibrant city of São Carlos that hosts campuses of the University of São Paulo (USP) and the Federal University of São Carlos (UFSCAR).

This volume contains the papers presented at ENIAC 2014. Papers were submitted in Portuguese and English and reviewed by at least 2, and on the average 3, program committee members. A total number of 97 papers will be presented in ENIAC, covering a wide range of topics in Artificial and Computational Intelligence. The papers in this proceedings are available online as a BDBComp proceedings via the site:
http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Evento?id=739

A selected number of papers will be selected to appear in special issues of the Progress in Artificial Intelligence Journal and the Advances in Distributed Computing and Artificial Intelligence Journal.

We would like to express our sincere thanks to all Program Committee members and to all reviewers for their cooperation in the reviewing process. We very much appreciated their hard work, which guaranteed the high quality of the technical program. We would like to extend our thanks to the Local Chairs and the volunteers who run several aspects of the organization and exhibition of this conference. Special thanks go to our Invited Speakers and to all of you who came to São Carlos to present and share your own work. Finally, we are most grateful for the sponsorship and support of CNPq, CAPES and FAPESP, as well as of Bloomberg, IBM Research and TERADATA.

September 1, 2014

Paulo Santos (FEI)
Ricardo Prudencio (UFPE)

BRACIS ENIAC 2014.

Program Co-Chairs
## Table of Contents

Ampliando o Perfil do Usuário para um Sistema de Recomendação de Nomes Próprios .......................... 1
  *Rafael Glauber and Angelo Loula*

A Trust-based -ATMS for the Future Internet .................. 7
  *Mondi Ravi, Yves Demazeau and Fano Ramparany*

Reconhecimento de Padrões Aplicados a Comentários de Fóruns Educacionais ................................. 13
  *Vitor Rolim, Filipe Rolim and Rafael Ferreira*

Deteccão Automática de Spammers em Redes Sociais ................ 19
  *Jhony M. Campanha, Johannes V. Lochter and Tiago A. Almeida*

Abordagem Semi-supervisionada para Rotulação de Dados .................. 25
  *Bruno Lima, Vinicius Machado and Rodrigo Veras*

Keyword Identification in Structured Reports in the Presence of Ambiguity .................. 30
  *Kuruvilla Joseph Abraham, Flavia Pena Nicolas, Amanda Da Rocha Reis, Ivan Torres Pisa and Evandro Eduardo Seron Ruiz*

Uma abordagem para a escolha do melhor método de seleção de instâncias usando meta-aprendizagem .................. 36
  *Shayane Moura, Marcelo Bassani Freitas, Halisson Cardoso and George Darmiton*

Normalização Textual e Indexação Semântica Aplicadas na Filtragem de SMS Spam .............................. 41
  *Tiago P. Silva, Igor Santos, Tiago A. Almeida and José M. G. Hidalgo*

Análise de Múltiplos Sentimentos em Textos Baseada em Lexicon ........ 47
  *Igor Barbosa, Ygor Diniz, Andy Gajadhar, Yuri Malheiros and Andrei Formiga*

On the data classification using complex network entropy .................. 53
  *Filipe A. Neto and Liang Zhao*

Uma arquitetura para combinação de classificadores otimizada por métodos de poda com aplicação em credit scoring .................. 59
  *Luiz Vieira E Silva Filho and George D.C. Cavalcanti*

Evolução Diferencial Aplicada ao Problema de Despacho Econômico de Energia Elétrica .................. 67
  *Lucas Prestes, Richard Gonçalves, Josiel Kuk, Sandra Venske and Carolina Almeida*
Classificação Automática de Gêneros Musicais Latinos Utilizando Sistemas Fuzzy

João Marcelo Bernardi, Marcos Henrique de Andrade, Glaucia Bressan and Carlos Silla Júnior

Geração de Regras Fuzzy para Classificação pelo Algoritmo OGP

Carlos Magno Valle, Adriano Soares Koshiyama, Ricardo Tanscheit and Marley Vellasco

Criminal Hot Spot Detection using Formal Concept Analysis and Clustering Algorithms

Adriana M. G. Farias, Marcos Cintra, Angélica Félix de Castro and Danniell Cavalcante Lopes

Técnicas de agrupamento aplicadas em culturas para suporte à decisão em agricultura de precisão.

Mariana Pereira, Kemilly Garcia, Claudio Ronchi and Murilo Naldi

Informação de Fisher no modelo de Potts definido em grafos para classificação supervisionada

Alan Hiraga and Alexandre Levada

RCMDE-GSD: Building Global Hierarchical Classifiers Using Differential Evolution for Predicting Gene Ontology Terms

Rafael Abud Menezes and Júlio Cesar Nievola

Problema de Alocação de Berços em Terminais de Contêineres através de Algoritmos Genéticos

Adriane Serapiao and Larissa Cristina Moraes

Um Modelo de Suporte à Comunicação de Agentes para tratamento de informação imprecisa com ontologia: integrando aos ambientes Jason e Jade

Fabio Aiub Sperotto and Diana Francisca Adamatti

Incorporating Object Features in Collaborative Argumentation-Based Negotiation Agents

Pablo Pidotti, Ana Casali and Carlos Chesñevar

Avaliação de Técnicas de Regressão para Predição de Fluxo de Veículos na Cidade de Belo Horizonte

Aline Xavier Fidêncio and Luiz Merschmann

Query Rules Study on Active Semi-Supervised Learning using Particle Competition and Cooperation

Fabricio Breve
Rede de Relacionamentos Brasileira de Inteligência Artificial e Computacional ......................................................... 141
   Luciano Antonio Digiampietri, Sarajane Peres and Leandro Silva

Computational Intelligence: Trailing Mendeleyev’s Footsteps ................. 147
   Maurício Ruv Lemes, Arnaldo Dal Pino Júnior and Kallaran Cavalcante Barros

Abordagem de Classificação baseada em Comitê para Alinhamento de Ontologias .............................................................. 152
   Vinicius Lopes, Fernanda Baião and Kate Revoredo

Investigando a Utilização de Abordagens Compactas nas Estratégias de Evolução ................................................................. 158
   Anderson Sergio, Sidartha Carvalho and Marco Rego

Improving Data Mining Results by taking Advantage of the Data Warehouse Dimensions: A Case Study in Outlier Detection ................. 164
   Mohammad Nozari Zarmehri and Carlos Soares

Hyperlab: A Java framework for the creation and management of hyper-heuristics and problem suites ........................................ 170
   Kamila M. Galvani and Fernando J. Von Zuben

Identificação Automática de Gêneros das Mensagens em Fóruns de Discussões do AVA ................................................................. 176
   Fabrício Guimarães and Ahmed Esmin

Modelos de Ação com Pós-Condição ............................................. 182
   Isaque Lima and Mario Benevides

Reasoning on Ontology Version Space with Temporal Logics ..................... 188
   Iuri Fernandes Queiroz, Luis Henrique Bustamante, Ana Teresa Martins and João Alcântara

Complex networks for aiding online discussion forum evaluation ............... 194
   Fabiano Berardo de Sousa, Elenise Maria Araújo and Jose Dutra Oliveira Neto

Viabilidade do Aprendizado Ativo em Máquinas Extremas ...................... 200
   Davi Santos and Andre Carvalho

A local decision making cellular automata-based path-planning ............... 207
   Gina Oliveira and Giordano Ferreira

Improving Local Gaussian Process For Real-Time Online Regression .......... 213
   Renato De P. Pereira, Paulo M. Engel, Thiago F. Rodrigues and Cassio Felipe F. De Oliveira
Multi-label Classification of Music into Genres .......................... 219
Vitor Da Silva and Ana T. Winck

TweeProfiles: Detection of spatio-temporal patterns on Twitter .......... 224
Tiago Cunha, Carlos Soares and Eduarda Mendes Rodrigues

Detecting Changes in 3D Maps using Gaussian distribution .......... 230
Sidnei Silva Filho, Paulo Drews-Jr and Silvia Botelho

A Robust dual layer framework solution to the Road Sign Problem ...... 236
Cassio Oliveira and Paulo M. Engel

Bioinspired Data Mining Algorithms: An Approach for Data Clustering
with Swarm Intelligence Algorithms ...................................... 241
Felipe Bonon Gonçalves, Guilherme Sanchez Corrêa, Adriane Beatriz
Souza Scrupião and Veronica Oliveira de Carvalho

Aplicação de redes neurais artificiais no estudo simulado da degradação
facultativa de detergentes ................................................. 249
Pierre Prado, Amanda Prandini, Iolanda Duarte and Antonio Martins

On the Checking of Indirect Normative Conflicts ....................... 254
Jean Zahn and Viviane Silva

A Robust and Regularized Extreme Learning Machine .................. 260
Ananda Freire and Guilherme Barreto

Probabilistic Ontologies Incremental Modeling Using UnBBayes ........ 266
Laécio L. Santos, Rommel Carvalho, Marcelo Ladeira and Li Weigang

An Agent-Based Metaheuristic Approach applied to Vehicle Routing
Problem with Time-Windows .............................................. 272
Maria Amélia Lopes Silva, Sergio Ricardo De Souza, Sabrina Oliveira
and Marcone Jamilson Freitas Souza

Sistema Nebuloso Preditivo para Manutenção de Transformadores de
Potência em Sistema de Medição de Gases Dissolvidos em Óleo Mineral
Isolante ................................................................. 278
Adriane Scrupião and Celso Modesto Jr.

Desenvolvimento de um Software para Detecção Automática de Tópicos
em Documentos Textuais Baseada em Taxonomia ..................... 285
Patrick Silva, Elvio Silva and Christian Freitas

A Novel Process Meta-model for Developing Automatic Speech
Recognition Systems ...................................................... 293
Gabriel Araujo and Hendrik Macedo
A comparison study of classifier algorithms: A submersible motor pump conditions in offshore oil exploration application .......................... 299
Alexandre Rodrigues, Flavio Miguel Varejão and Marcos Pellegrini Ribeiro

A proposed SLAM for underwater vehicles ................................. 305
Felipe Guth, Silvia Botelho, Luan Silveira, Paulo Drews-Jr and Matheus Machado

Aperfeicoamento do Mapeador de Teses e Dissertações da UFPE ........ 311
Ubiracy Dos Santos Rego Junior, Teresa Bernarda Ludermir and Renato Fernandes Correa

Semantic Web data representation in BDI Agents .......................... 318
Diogo De Campos and Ricardo Azambuja Silveira

Modelos de Regressão para a Previsão de Séries Temporais por meio do Algoritmo kNN-TSP ......................................................... 323
Carlos André Ferrer, André Gustavo Mallettza and William Zalewski

Esquema de Alocação de Blocos de Recursos com Garantia de QoS Baseado em Lógica Fuzzy para Redes LTE .............................. 330
Diego Cruz Abraão and Flávio Henrique Teles Vieira

Metodologia para Avaliar Técnicas de Redução de Protótipos:
Protótipos Gerados versus Protótipos Selecionados ......................... 336
Luciano De Santana Pereira and George D.C. Cavalcanti

Aplicação do Algoritmo ACO-HH para o problema de cobertura de conjuntos ................................................................. 342
Alexandre S. Ferreira, Aurora T. R. Pozo and Richard Aderbal Gonçalves

Metodologia baseada em redes complexas para análise das votações de deputados brasileiros ...................................................... 347
Fabiano Berardo de Sousa and Liang Zhao

Resolução Para o Problema n-Rainhas Utilizando ACO .................... 353
Carolina Moreira and Aurora Pozo

Um estudo sobre Otimização por Partículas aplicado ao problema de roteamento de veículos com demandas estocásticas ................... 359
Vinícius Renan de Carvalho and Aurora Pozo

Configuração automática de parâmetros: um estudo de caso ............. 365
Gian Frutsche and Aurora Pozo

New Approach to Detect the Political Opinion in Tweets .................. 371
Diaa Ezzeddine, Fabien Rico and Djamel A. Zighed

Classificação de padrões robusta com redes Adaline modificadas ........ 377
César Lincoln Mattos, José Daniel Santos and Guilherme Barreto
Segmentação Espacial Não Uniforme Aplicada ao Reconhecimento de
Gênero e Expressões Faciais ........................................ 383
   Vagner Amaral, Gilson Giraldi and Carlos Thomaz

Who is their mother?: A classification work to get answers over
registration people databases ...................................... 389
   Gustavo C. G. Van Erven, Rommel Novaes Carvalho, Maristela Holanda,
   Marcelo Ladeira, Henrique Rocha and Gilson Mendes

Estudo do Impacto de um Corpus Desbalanceado na Identificação de
Emoções em Textos .................................................. 394
   Lohann Paterno Coutinho Ferreira, Mariza Miola Dosciatti and Emerson Cabrera Paraíso

Bengala Inteligente Neural baseada em Aprendizagem por Reforço para
Deficientes Visuais .................................................. 401
   Franciele A. S. Alves, Alexandre M. M. Neumann and Maury M. Gouveia Jr.

Alinhamento múltiplo de sequências utilizando otimização dialética...... 407
   Rodrigo Gomes De Souza, Antônio Luiz Vieira Da Silva Júnior, Ricardo Yara and Wellington Pinheiro Dos Santos

A comparison of the effect of feature selection and balancing strategies
upon the sentiment classification of Portuguese News Stories............. 413
   Brett Drury and Alneu de Andrade Lopes

Aplicação do algoritmo sarsa à coleta de lixo - Avaliação de parâmetros . 418
   Darlinton Prauchner, Rogério Martins and Edson Padoin

Utilizando Reconhecimento Semântico de Objetos na Formulação de
Comportamentos Adaptativos na Navegação de Robôs Móveis ............ 424
   Jurasildo Reinaldo, Rosiery Maia and Anderson Souza

Aplicando técnicas de aprendizado de máquina em planejamento
probabilístico ....................................................... 430
   Jean L. Sousa and Carlos Roberto Lopes

Uma abordagem de alinhamento múltiplo de sequências utilizando
evolução diferencial .................................................. 436
   Antônio Luiz Vieira Da Silva Júnior, Rodrigo Gomes De Souza, Ricardo Yara and Wellington Pinheiro Dos Santos

Agente negociador baseado em técnicas fuzzy .......................... 441
   Miriam Mariela Morveli Espinoza, Myriam Regattieri Delgado and César A. Tacla

Machine Learning And Adaptive Morphological Operators .............. 447
   Magno Almeida Filho, Francisco Silva and Arthur Braga
ClusterOSS: a new undersampling method for imbalanced learning
Victor H Barella, Eduardo P Costa and André C P L F Carvalho

Application of text mining techniques for classification of documents: a study of automation of complaints screening in a Brazilian Federal Agency
Patrícia Andrade, Marcelo Ladeira, Rommel Carvalho, Henrique Rocha and Gilson Libório

Evolução Diferencial com Múltiplos Vetores Experimentais Aplicada ao Problema do Despacho Econômico de Energia Elétrica
Erick Cesaro, Gustavo Czaikoski, Richard Goncalves, Carolina Almeida, Sandra Venske and Josiel Kuk

Inteligência de Enxames na Otimização da Densidade de Redes de Sensores Sem Fio
Carlos Henrique Drummond, Paulo Roberto Ferreira and Lisane Brisolara

Aprendizado por reforço em lote para o problema de tomada de decisão em processos de venda
Denis Antonio Lacerda and Leliane Nunes de Barros

Otimização multiobjetiva com sistemas imunológicos artificiais e operador de supressão
Ricardo de Carvalho Destro and Reinaldo A. C. Bianchi

Aprendizagem por Reforço com Rede Neural no Desenvolvimento de Jogos Digitais
Edival Assis and Maury Gouveia

Construção automática de algoritmos de indução de árvores de decisão: uma abordagem multiobjetiva
Melis Silva, Rodrigo Barros and Márcio Basgalupp

Problema de Intervenção em Redes Gênicas Modelado como um Processo de Decisão Markoviano Fatorado
Fabio Tisovec, Leliane de Barros, Karina Delgado and Ronaldo Hashimoto

Analysing some t-norm-based generalizations of the Choquet Integral for different fuzzy measures with an application to fuzzy rule-based classification systems
Giancarlo Lucca, Rogério R. De Vargas, Graçaliz Dimura, José Antonio Sanz Delgado, Humberto Bustince and Benjamin Bedregal

Solving Sokoban Optimally using Pattern Databases for Deadlock Detection
André Grahl Pereira, Marcus Ritt and Luciana Salete Buriol
Investigaçãode técnicas de otimização baseada em população em conjunto com técnicas de busca local para classificação de dados

Gustavo Custódio and Debora Medeiros

Algoritmo Fuzzy para Controle de Tráfego de Rede Baseado em Modelo Multifractal Adaptativo e Funções de Base Orthonormais

Alisson Cardoso, Flávio Vieira and Diego Abrahão

ASDP: um processo para Análise de Sentimento em Debates Polarizados

Francisco Assis Ricarte Neto and Flávia Barros

Real-time Sentiment Analysis in Social Media Streams: The 2013 Confederation Cup Case

Paulo Cavalin, Maira Gatti, Cícero Nogueira Dos Santos and Claudio Pinhanez

Um Modelo Híbrido para Previsão de Produção de Petróleo

Francisca De Fátima Silva, Adrião Duarte Dória Neto, Paulo Sérgio Lucio and Eduardo Henrique Silveira de Araujo

Identificando o Assunto dos Documentos em Coleções Textuais Utilizando Termos Compuestos

Fabiano Fernandes Dos Santos, Veronica Oliveira de Carvalho and Solange Oliveira Rezende

Multi-objective Evolutionary Membership Functions Tuning as a Post-processing Task in the Generation of Fuzzy Classification Systems

Edward Hinojosa Cardenas and Heloisa Camargo

Análise de uma Estratégia de Coordenação Bio-inspirada de Múltiplos Robôs para a Tarefa de Vigilância em Ambientes Desconhecidos

Rodrigo Calvo, Janderson R. De Oliveira, Maurício Figueiredo, Roseli Romero and Ademir Constantino

Uma abordagem de poda para Máquinas de Aprendizado Extremo via Algoritmos Genéticos

Alisson S. C. Alencar and Ajalmar Régo Da Rocha Neto

Descoberta de Conhecimento com Auxílio da Inteligência Humana: um estudo de caso para dobramento de proteínas

Renan Luz, Diana Franciscia Adamatti and Adriano Werhli

Bi-objective Worker Assignment in the bases of StarCraft

Caio Freitas Oliveira, Elizabeth Elizabeth Ferreira Gouvea Goldbary and Marco Cesar Goldbary
ClusterOSS: a new undersampling method for imbalanced learning

Victor H Barella, Eduardo P Costa, and André C P L F Carvalho,

Abstract—A dataset is said to be imbalanced when its classes are disproportionately represented in terms of the number of instances they contain. This problem is common in applications such as medical diagnosis of rare diseases, detection of fraudulent calls, signature recognition. In this paper we propose an alternative method for imbalanced learning, which balances the dataset using an undersampling strategy. We show that ClusterOSS outperforms OSS, which is the method ClusterOSS is based on. Moreover, we show that the results can be further improved by combining ClusterOSS with random oversampling.

Keywords—Imbalanced data, classification, sampling, clustering.

1 INTRODUCTION

A dataset is said to be imbalanced when its classes are disproportionately represented in terms of the number of instances they contain. For example, in a problem about a rare disease, the number of cases of infected people is usually much lower than that of healthy people. Other examples of problems containing imbalanced data are: detection of fraudulent calls [1]; detection of fraudulent credit cards transactions [2]; and signature recognition [3].

Traditional machine learning methods usually yield unsatisfactory results in problems of this nature. While they present good results for the majority classes, they perform very poorly w.r.t. the minority classes. The main problem with this is that, in imbalanced learning, the minority classes are usually the classes we are interested in. For this reason, many methods have been proposed, as we discuss in Section 2. One popular method in this context is OSS (One-sided Selection) [4], which artificially balance the dataset by disregarding instances of the majority class which look to be redundant.

In this paper we propose an alternative method for imbalanced learning in the context of binary classification; the proposed method is based on OSS. We show that our method yields better results than OSS. As an additional contribution of this paper, we present an empirical comparison among six methods, including OSS and our proposed method.

The remaining of this paper is organized as follows. In section 2 we discuss related work. In section 3 we present our proposed method. We present our experimental results in Section 4. We conclude in Section 5.

2 RELATED WORK

There are two main general approaches for classification problems involving imbalanced data: (1) pre-processing the data in order to make it more balanced; and (2) development of algorithms able to handle imbalanced data. In this paper we focus on the former.

Pre-processing methods can be categorized into two groups: undersampling and oversampling methods. Undersampling methods make the data more balanced by removing instances of the majority class, while oversampling methods do that by inserting instances in the minority class. Both undersampling and oversampling can be done randomly or according to an informative strategy. Next we discuss the main pre-processing strategies.
2.1 Random undersampling

In the random undersampling, instances of the majority class are removed at random until a more balanced class distribution is reached.

2.2 Random oversampling

In the random oversampling, instances of the minority class are replicated at random until a more balanced class distribution is reached.

2.3 SMOTE

SMOTE (Synthetic Minority Oversampling Technique) [5] is an oversampling technique that creates artificial data by interpolation, as follows. At each iteration, SMOTE selects an instance \( x \) at random in the minority class and then it looks for the \( k \) nearest neighbors of \( x \). SMOTE then selects one of the neighbors \( z \) at random and creates a new instance which is a combination of \( x \) and \( z \). This step is repeated until a more balanced distribution of instances is reached.

2.4 CBO

CBO (Cluster-Based Oversampling) [6] is an oversampling technique that takes into account both inter- and intra-class imbalance. Intra-class imbalance occurs when there is a disproportion w.r.t. the instances which form subsets inside a class.

This technique starts by performing a clustering procedure in the instances belonging to the majority class and the instances belonging to the minority class, separately; this step will generate two sets of clusters - one for each class. Next, CBO applies random oversampling to all clusters belonging to the majority class with exception of the largest one. In the end, each cluster of the majority class should have the same number of instances as the largest one. Finally, oversampling is applied to all clusters belonging to the minority class such that in the end (1) the total number of instances in the minority class equals the total number of instances in the majority class after the oversampling, and (2) each cluster in the minority class has the same number of instances.

2.5 OSS

OSS (One-sided Selection) [4] is an undersampling technique that keeps only the most representative instances of the majority class. To select those instances, OSS first chooses one instance \( x \) of the majority class at random. Then, using the instances of the minority class and \( x \) as training data, OSS uses the k-Nearest Neighbors (KNN) algorithm with \( k=1 \) to classify the remaining instances of the majority class. The correctly classified instances are then excluded of the majority class because they are considered redundant. Thus, after the undersampling, the majority class will contain only the instances which were incorrectly classified and \( x \). Finally, OSS uses a data cleaning technique to remove borderline and noisy instances, which is, originally, Tomek Links [7].

3 PROPOSED METHOD

This section introduces our proposed method, which is based on the OSS strategy. We first motivate our method, by pointing out situations in which OSS might not work well. Then, we introduce our method, called ClusterOSS.

3.1 Motivation

OSS assumes that is enough to choose only one majority instance at random to start the undersampling process. However, the final result of the undersampling method will depend on that random choice. More importantly, OSS does not explicitly take into account the fact that there might exist subsets inside the majority class (as CBO does, for example), and that the undersampling might not work equally well in all those subsets, given the random start.

Consider, for example, the dataset displayed in Fig. 1. In the figure, the instances of the majority class are represented by circles, while the instances of the minority class are represented by triangles. Note that the majority class is divided into two subsets - one to each side of the minority class. Fig. 2.a shows the randomly selected majority instance together with the minority instances. Fig. 2.b shows the resulting dataset given by OSS.

Note that the effect of the undersampling process is limited by the fact the majority class contains two subsets and by the choice of the
instance to start the process (which is far from the minority class in the feature space). Most of the instances of the majority class are kept after the undersampling process, resulting in a dataset which is still very imbalanced.

Suppose now the instance would have been chosen in the center of the cluster at the right of the figure. In this case, the undersampling would have worked well for that cluster, but would have had little (or no) effect on the other subset of the majority class.

Next, we present an alternative method that avoid these situations.

3.2 ClusterOSS

Our proposed method (ClusterOSS) is an adaptation of the strategy used by OSS. Before describing the ClusterOSS algorithm, we point out the two main differences of ClusterOSS w.r.t. OSS.

The first difference is that ClusterOSS can start the undersampling process from more than one instance. This, in itself, already tackles the drawback of OSS that the quality results is strongly dependent on the choice of that one instance chosen to start the undersampling.

The second difference is that we do not start the undersampling process at random. Instead, we define how many and which instances will be chosen to start that process. More specifically, we look for subsets in the majority class, by applying a clustering procedure. Then we choose the instance at the center of each subset to be one of the instances which will start the undersampling. By doing this, we enhance the effectiveness of undersampling, since the undersampling will start from points in distinct regions in the feature space.

Algorithm

The ClusterOSS algorithm is showed in Fig. 3.a in the form of pseudocode. In the beginning of the algorithm we use a clustering procedure (e.g., k-means) to cluster the instances belonging to the majority class. Then, for each cluster, we use the closest instances to the center. These instances are used to start the undersampling process, which is identical to OSS. Finally, as in OSS, we use the data cleaning technique Tomek Links (Fig. 3.b) to remove borderline and noisy instances. Basically, this technique removes every instance $z$ from the majority class for which (1) its closest neighbor $z'$ is an instance of the minority class, and (2) the closest neighbor of $z'$ is also $z$.

Example

We illustrate how ClusterOSS works using the same dataset we showed in Fig. 1. Fig. 4.a shows the selected majority instances (each of them being in the center of the subsets identified by k-means) together with all minority instances. Fig. 4.b shows the resulting dataset given by ClusterOSS.

Note that ClusterOSS is able to obtain a more balanced dataset (Fig. 4.b) than that obtained by OSS (Fig. 2.b). The original dataset has a proportion of 1:40 (minority class:majority class), while the proportions of the resulting datasets are approximately 1:30 and 1:5 for OSS e ClusterOSS, respectively. It is important to mention that both strategies reduce the majority class in distant regions from the minority class. It is the Tomek Links step that acts on the closer region to the minority class.

4 Experiments

We present an empirical evaluation of our method. The main goal of it is to verify whether the alternative undersampling strategy used by ClusterOSS yields better results than those
4.1 Experimental Settings

We implemented ClusterOSS with the clustering method $k$-means, and we determine the number of clusters by the average silhouette of the training set. For $k$-means we consider squared Euclidean distance as proximity measure, 10 as maximum number of iteration and 1 initial configuration. To obtain the average silhouette we consider the Euclidean distance as proximity measure. For CBO we use the same strategy. We combined SMOTE with random undersampling as suggested by its creators. We use random oversampling and random undersampling with final proportion of 1:1. We use SMOTE with parameters such it increases the minority class in 200% and decreases the majority class such that the final number of instances in the minority class is 75% of the final number of instances in the majority class. To test the quality of the pre-processing of the data, we apply three different classification algorithms - KNN ($k=3$), C5.0 and SVM - to the resulting dataset given by each method.

The evaluation was performed on 10 datasets, which are showed in Table 1; the table contains the name, number of attributes (including the class attribute), number of instances and proportion ratio of the datasets.

We obtained the datasets Vowel, Haberman, Pima Diabetes and Yeast from the UCI repository[8], and Cleveland, Poker and Vehicle from the Keel repository[9]. Vowel, Yeast, Cleveland, Poker and Vehicle are originally multiclass problems and were turned into binary problems by choosing a specific class as the positive one and making the following relation of positive classes x negative classes: 0 x rest, 4 x rest, 0 x 4, 8 x 6 and 2 x rest, respectively.

The three artificial datasets were created using a normal distribution for each class (or for each subset of the class, when the class is divided in more than one subset). The artificial datasets are used to assay the performance of the techniques in different situations. They are all binary problems and have three attributes(two are numeric and one is the class). They are plotted in the Figure 5, where the 'X' are instances from the majority class and the...
Fig. 5: Artificial Datasets. Top-left: Artificial dataset (a). Top-right: Artificial dataset (b). Bottom: Artificial dataset (c).

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Attributes</th>
<th># Examples</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial (a)</td>
<td>3</td>
<td>410</td>
<td>1 : 40</td>
</tr>
<tr>
<td>Artificial (b)</td>
<td>3</td>
<td>510</td>
<td>1 : 50</td>
</tr>
<tr>
<td>Artificial (c)</td>
<td>3</td>
<td>520</td>
<td>1 : 25</td>
</tr>
<tr>
<td>Vowel0</td>
<td>11</td>
<td>990</td>
<td>1 : 10</td>
</tr>
<tr>
<td>Haberman</td>
<td>4</td>
<td>306</td>
<td>1 : 3</td>
</tr>
<tr>
<td>Yeast4</td>
<td>8</td>
<td>1479</td>
<td>1 : 28</td>
</tr>
<tr>
<td>Pima Diabetes</td>
<td>9</td>
<td>768</td>
<td>1 : 1.86</td>
</tr>
<tr>
<td>Cleveland0x4</td>
<td>14</td>
<td>173</td>
<td>1 : 12.31</td>
</tr>
<tr>
<td>Poker8x6</td>
<td>11</td>
<td>1477</td>
<td>1 : 85.88</td>
</tr>
<tr>
<td>Vehicle2</td>
<td>19</td>
<td>846</td>
<td>1 : 2.88</td>
</tr>
</tbody>
</table>

**TABLE 1: Dataset Information**

circles are from the minority class. The dataset (a) has two majority regions and one minority region between them. The dataset (b) has one majority and one minority region completely overlapped. The dataset (c) has a majority region rounded by two minority regions.

We perform the experiments with stratified 5-fold cross validation; we do it 100 times. We choose the stratified variant to keep the class distribution in each fold, and we choose 5 folds to avoid situations in which there is too few examples in the minority class to be able to apply the classification methods.

We used the following evaluation measures: Positive Accuracy \( \frac{TP}{TP+FN} \), Negative Accuracy \( \frac{TN}{TN+FP} \), Geometric Mean of Accuracies \( \sqrt{Pos.Accuracy \times Neg.Accuracy} \) and the Area Under the Roc Curve (AUC).

<table>
<thead>
<tr>
<th>Measure</th>
<th># OSS victories</th>
<th># ClusterOSS victories</th>
<th>Ties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Accuracy</td>
<td>4</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>Negative Accuracy</td>
<td>13</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Geometric Mean</td>
<td>5</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>AUC</td>
<td>9</td>
<td>21</td>
<td>0</td>
</tr>
</tbody>
</table>

**TABLE 2: OSS x ClusterOSS**

4.2 Results and Discussion

First, we compare OSS and ClusterOSS. Table 2 shows the number of wins and ties for them. For each line of the table, we show the results for one of the evaluation measures, considering the results for the combination of the 10 datasets and the 3 classification methods. The results show that ClusterOSS yields better results for the positive class (minority class), which is the class of interest. The results for the negative class are comparable, with a slight advantage for OSS. This shows a trade-off between the performance in the positive and negative classes. However, when we consider measures that evaluate the performance on both positive and negative classes - geometric mean and AUC - ClusterOSS outperforms OSS.

We now perform a comparative analysis of all 8 methods considered in our experiments. To summarise the results, we count the number of victories for each pre-processing technique in each dataset and in each classification algorithm for different measure performances. The results are shown in Fig. 6 which represents this rank. The dark bar represents the number of victories a technique perform compared to the others, and the white bar represents how many times a technique is in the top 3 performances.

First of all, we can see that the dataset without pre-processing yields the best results for the negative class, but very poor results for the positive class. This is expected, since the imbalance of the datasets leads to a bias w.r.t. the negative class.

Random undersampling presents an opposite behaviour. While, it gives the best performance w.r.t. the positive accuracy, it presents a poor performance w.r.t. the negative accuracy when compared to the other methods. Because of this difference in the results, random undersampling is outperformed by other methods (e.g., SMOTE and ClusterOSS followed by ran-
dom oversampling) w.r.t. AUC and the geometric mean.

Even though we saw that ClusterOSS outperforms OSS, the comparative analysis with all methods shows that ClusterOSS does not rank among the best methods. However, when we combine ClusterOSS and random oversampling, the results are compared to those of SMOTE; these 2 methods being the best ones in a general analysis of the measures.

Now, we have a closer look at the results of ClusterOSS followed by random oversampling, and SMOTE. We show the comparative analysis of these 2 methods in Table 3. We can see that while SMOTE performs better w.r.t. the positive accuracy, ClusterOSS followed by random oversampling performs better w.r.t the negative accuracy. For the other 2 measures the results are comparable.

5 Conclusions and Future Work

In this paper, we introduced a new under-sampling method to pre-process imbalanced datasets. Our method, which we call ClusterOSS, outperforms OSS, which is the method ClusterOSS is based on. Moreover, we showed that when we combine ClusterOSS with random oversampling, the results are comparable with those of the state-of-the-art SMOTE.

As future work, we will investigate why SMOTE performs better than ClusterOSS with random oversampling regarding the positive class, in order to see possible directions in which we can improve our results.

Acknowledgments

The authors would like to thank the Brazilian research agencies FAPESP, CAPES and CNPq for financially supporting this work.

References


Table 3: SMOTE x ClusterOSS followed by random oversampling

<table>
<thead>
<tr>
<th>Measure</th>
<th># SMOTE victories</th>
<th># ClusterOSS + random oversampling victories</th>
<th>Ties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Accuracy</td>
<td>18</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Negative Accuracy</td>
<td>9</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>Geometric Mean</td>
<td>16</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>AUC</td>
<td>15</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>