Minerando dados na Infraestrutura da Internet: analisando a participação em eventos do IETF

Julião Braga[1], **Leandro Augusto da Silva**[1], **Nizam Omar**[1]

[1]Departamento de Engenharia Elétrica - Universidades Presbiteriana Mackenzie
São Paulo - SP - Brasil

juliao@braga.eti.br,leandroaugusto.silva@mackenzie.br,nizam.omar@mackenzie.br

Abstract. *This work aims to analyze the participation of the regional Internet registries in the main events of the IETF through techniques of data mining with emphasis on scatter plots and the results of the confusion matrices produced by classification trees, and optimization capabilities.*

Resumo. *Este trabalho tem como objetivo analisar a participação dos registros regionais da Internet nos principais eventos do IETF através de técnicas de mineração de dados com ênfase nos diagramas de espalhamento e nos resultados das matrizes de confusão produzidas pelas árvores de classificação, com as respectivas alternativas de otimização.*

1. Introdução

Neste trabalho utilizam-se dados referentes às inscrições aos encontros anuais realizados pelo *Internet Engineering Task Force*[1] (IETF), que somente começaram a ser disponibilizados a partir do encontro 72, realizado em 2008[2]. A Figura 1 exibe um trecho da tabela de inscrições, do encontro IETF 89[3].

Last Name	First Name	Organization	ISO 3166 Code	On-Site	Profile
Aas	Joshua	Mozilla Corporation	US	Yes	
Abdullah	Musab	Bahrain Telecommunications Regulatory Authority	BH	No	
Abley	Joseph	Dyn, Inc.	CA	Yes	YES
Aboba	Bernard	Microsoft Corporation	US	Yes	YES
Abrahamsson	Mikael	Deutsche Telekom	SE	Yes	
Accettura	Nicola	Politecnico di Bari	IT	Yes	

Figura 1. Extrato das inscrições no IETF 89. (Fonte: IETF)

A motivação principal em usar tais bases repousa no interesse dos autores em desenvolver trabalhos relacionados com aplicações de Web Semântica, na Infraestrutura da Internet[4]. É de interesse investigar a participação muito pequena, de pessoas da América Latina, em particular, do Brasil, nos eventos do IETF, como se pode verificar no gráfico da Figura 2, em uma comparação com a participação global.

Anteriormente a esta motivação houve a preocupação em divulgar[5] as iniciativas do IETF por força de uma bolsa recebida, por um dos autores, para participar do encontro 86 (março de 2013), em Orlando, FL.

[1]http://www.ietf.org
[2]http://www.ietf.org/meeting/past.html
[3]https://www.ietf.org/registration/ietf89/attendance.py
[4]http://ws.org.br
[5]Tal divulgação ocorreu através de alguns artigos no blogue Infraestrutura da Internet, http:

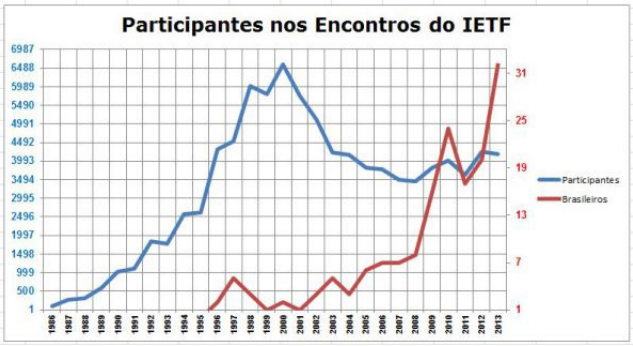


Figura 2. Participação brasileira nos encontros do IETF[6]

.

Tais iniciativas começaram no final de 2012 e continuaram em um processo de divulgação intensiva durante todo o ano de 2013. Notou-se um aumento na participação, enquanto que novidades como os eventos[7] pré IETF eclodiram da América Latina e, mais recentemente, o edital público do CGI.br ofertando bolsas para pesquisas e participação nas reuniões do IETF por três anos consecutivos, disponibilizadas para empresas e universidades[8].

Pelas observações acima se percebeu a necessidade de haver continuidade na prospecção dos dados, principalmente avaliando as possibilidades de novos resultados.

Da Figura 1, o interesse foi unicamente sobre o atributo *ISO_3166 Code*, um código de dois caracteres que identifica o respectivo país do inscrito. Este código é conhecido como *country code Top Level Domain* (ccTLD) e representa o país no nome de domínio usado na Internet, principalmente, nos navegadores. Ele foi usado para o mapeamento da participação de algumas regiões do mundo. Por outro lado havia uma questão sobre se o local da realização do evento influenciaria na participação destas regiões. O IETF realiza seus encontros (três, por ano), em diversas cidades do mundo. As localidades onde os eventos foram realizados podem ser vistas na Figura 3.

Mineração de dados é a descoberta e extração (mineração) de conhecimento a partir de grandes bases de dados. [Fayyad et al. 1996] dizem que *o processo de descoberta de conhecimento é interativo e iterativo, enfatizando a aplicação de técnicas*, que envolvem as etapas de (a) Preparação ou pré-processamento de dados, (b) Mineração de dados e (c) Análise dos resultados.

A grande maioria dos autores concorda que a realidade da mineração de dados está alicerçada em dois fundamentos – o conhecimento sobre o domínio da aplicação e a experiência do minerador – para identificar as tarefas de mineração que podem

---

`//bit.ly/PjRv1h` e `http://bit.ly/16YY81a`, uma página dedicada, `http://ietf.protocolos.net.br/`, sobre o IETF, a tradução, para o português, do "The Tao of IETF", `http://www.ietf.org/tao-translations.html` e da aprovação de recurso do CGI.br para a publicação do "Livro do IETF, a ser distribuído na América Latina, entre outras iniciativas.

[6] `http://ii.blog.br/2013/08/29/tip/`

[7] `http://iwietf.lacnog.org`

[8] `http://www.cgi.br/resolucoes/documento/2013/047`

[9] `http://ws.org.br/index.php/IETF_Meetings`

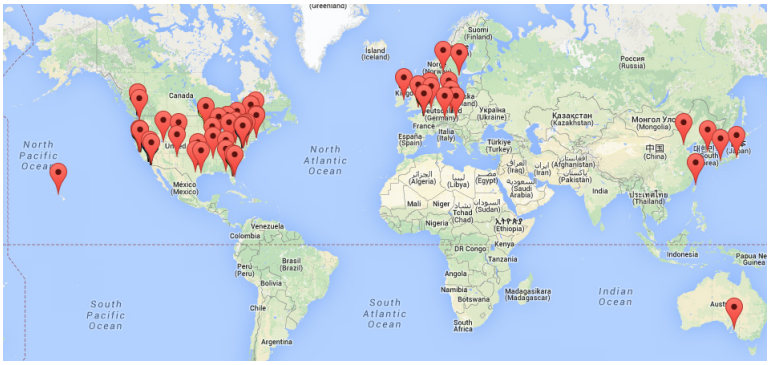


Figura 3. Localidades do encontros do IETF. Elaborado pelos autores[9].

ser aplicadas com o objetivo de consolidar o cenário de descoberta dos padrões na direção de resultados eficazes. O trabalho seguirá esta proposta usando diversas ferramentas e, em particular o programa *RapidMiner*[10].

O presente trabalho é dividido em seis seções, incluindo a Introdução. A segunda seção, Revisão de Conceitos trata de uma pequena descrição das etapas de mineração de dados. Nesta oportunidade serão apresentados as propostas de algoritmos utilizados pela mineração de dados e os recursos disponíveis para avaliação dos resultados. A terceira seção trata da Metodologia, onde serão caracterizadas as bases de dados e descritas as atividades de pré-processamento que foram necessárias para obtê-las, tanto externas quanto aquelas processadas no ambiente do *RapidMiner*. Na quarta seção, serão expostos os resultados e as respectivas discussões / avaliações da aplicação de uma tarefa de mineração de dados. A quinta seção trata das conclusões reproduzidas pela experiência obtida no uso dos algoritmos, a ferramenta utilizada (*RapidMiner*) e propostas de eventuais, e já existentes, incursões futuras. Finalmente, a sexta seção expõe as referências bibliográficas utilizadas.

2. Revisão de Conceitos

Quando se constroi um modelo durante o processo de aprendizagem, o objetivo é usá-lo para produzir novos dados, o mais próximo possível, da realidade. Aos dados sobre os quais o modelo é construído dá-se o nome de *conjunto de treinamento* e são reconhecidos como *dados visíveis*. Os novos conjuntos de dados, produzidos pelo modelo são chamado de dados *não-visíveis*. As técnicas envolvidas no processo de modelagem se preocupam com a eficiência do modelo escolhido, de tal forma que não ocorra uma generalização (transbordamento / *overfiting*) dos erros sobre os dados visíveis na direção dos dados não-visíveis (gerados pelo modelo), [Witten et al. 2011], [Chisholm 2013] e [Tan et al. 2009]. Por isto, algumas tarefas de mineração exigem o denominado *conjunto de teste*, representando os dados *não-visíveis*. Algumas tarefas podem necessitar de um terceiro conjunto de dados, o *conjunto de validação*, cuja presença aperfeiçoa o modelo (e não pode ser usado para testes). Todos estes três conjuntos de dados (*treinamento*, *teste* e *validação*) devem possuir exemplares de dados apropriados ao modelo. Tais técnicas e recursos estão amplamente discutidas no Capítulo 5 de [Witten et al. 2011], incluindo os indicadores de eficiência

[10]`http://www.rapid-i.com`

comumente utilizados.

2.1. Preparação ou Pré-processamento

Pré-processamento é um conjunto de técnicas que permitem adequar dados originais para uso na etapa da mineração de dados. O pré-processamento envolve a limpeza, a integração, a transformação, a redução, a discretização e a normalização de dados necessárias a tal adequação. Não é uma etapa distinta, pois o pré-processamento é usado, sistematicamente, durante todo o processo de descoberta de padrões, permitindo ao minerador aperfeiçoar o conhecimento adquirido sobre os dados, em resultados efetivos, durante as análises projetadas.

2.2. Mineração de dados

O trabalho da mineração de dados é dividido em tarefas de natureza *preditiva* ou *descritiva* e, as principais delas são: *associação*, *classificação*, *agrupamento* e *detecção de anomalias*. A *associação* tem como objetivo encontrar Regras de Associação entre itens que ocorrem simultâneamente. A *classificação* tem como objetivo o aprendizado de classificar um novo exemplar, a partir de exemplares originais, que contenham duas ou mais classes (atributos rotulados com valores discretos). A tarefa de *agrupamento* segmenta uma base de dados originais em subgrupos com características similares. A *deteção de anomalias* é uma tarefa que procura distinguir elementos com características distorcidas da grande maioria dos dados, [Witten et al. 2011].

Basicamente, os algoritmos inseridos nas tarefas de mineração são de dois tipos: *supervisionados* e *não supervisionados*. Os primeiros são aqueles sobre os quais o especialista em mineração de dados (o minerador) consegue intervir antes e durante a sua execução. Já os algoritmos *não supervisionados* são aqueles em que o minerador não consegue intervir.

2.3. Interpretação / Avaliação

Esta é a etapa final do processo de busca por padrões sobre os dados originais. Seguindo os objetivos definidos na etapa de avaliação e entendimento dos dados originais, o minerador, através de tarefas de mineração consegue enriquecer o conhecimento humano para tomadas de decisões.

3. Metodologia

3.1. As bases de dados do trabalho

A primeira preocupação foi reduzir as bases originais dos participantes no evento em cinco regiões da Terra, correspondentes à influência dos chamados *Regional Internet Registers*[11] (RIRs), organizações independentes e não governamentais autorizados pelo *Internet Corporation for Assigned Names and Numbers*[12] (ICANN), através de uma de suas organizações denominada *Internet Assigned Numbers Authority*[13] (IANA), para controlar os números e nomes usados na Infraestrutura da Internet. A grande maioria destes identificadores são padronizados pelo IETF. Tais divisões, ou regiões podem ser vistas na Figura 4.

[11] `https://www.iana.org/numbers`
[12] `http://www.icann.org/`
[13] `http://www.iana.org/`

Registry	Area Covered
AFRINIC	Africa Region
APNIC	Asia/Pacific Region
ARIN	North America Region
LACNIC	Latin America and some Caribbean Islands
RIPE NCC	Europe, the Middle East, and Central Asia

Figura 4. RIRs e respectivas regiões. Fonte: IANA.

Isto foi feito para cada um dos eventos, disponíveis no sítio do IETF. A primeira base de dados escolhida foi o resultado das inscrições do encontro 89 do IETF. Ela foi chamada de *BASE-1*, representada parcialmente na Figura 5 e disponível, integralmente, em http://ietf.protocolos.net.br/meetings/ietf89.php. Seus atributos são: o número de inscritos por país (*#*), o percentual de inscrições por país (*%*), o percentual de inscrições por RIR (*%RIR*), o RIR (*RIR*), o país (*Country*) e o código do país (*ccTLD*). Bases equivalentes de todos os encontros desde o 72 até o 88 encontram-se disponíveis na mesma URL acima substituindo o número 89, pelo respectivo número do evento.

Country	ccTLD	#	%	%RIR	RIR
Algeria	DZ	1	0.06	2.08	AFRINIC
Argentina	AR	1	0.06	4.55	LACNIC
Australia	AU	13	0.83	5.00	APNIC
Austria	AT	6	0.38	0.96	RIPE
Bahrain	BH	1	0.06	0.16	RIPE
Belgium	BE	28	1.78	4.47	RIPE
Benin	BJ	1	0.06	2.08	AFRINIC
Bolivia	BO	1	0.06	4.55	LACNIC
Botswana	BW	1	0.06	2.08	AFRINIC
Brazil	BR	7	0.45	31.82	LACNIC
Brunei Darussalam	BN	1	0.06	0.38	APNIC
Burkina Faso	BF	1	0.06	2.08	AFRINIC
Burundi	BI	1	0.06	2.08	AFRINIC
Cameroon	CM	1	0.06	2.08	AFRINIC
Canada	CA	35	2.23	5.70	ARIN

Figura 5. BASE-1 (parcial). Fonte: http://ietf.protocolos.net.br/meetings/ietf89.php

A segunda base de dados utilizada neste texto, *BASE-2*, contém o *número de inscrições* por evento, em cada país e um atributo no qual relaciona o país com a respectiva região. Na época em que a captura dos dados foi realizada, a *BASE-2* foi construída pelos eventos de número 72 a 85 ocorridos entre os anos de 2008 a 2012, vista na Figura 6.

			2008		2009			2010			2011			2012		
País	ccTLD	RIR	72	73	74	75	76	77	78	79	80	81	82	83	84	85
Afghanistan	AF	APNIC	0	0	0	0	0	0	1	1	0	0	0	1	0	1
Aland Islands	AX	RIPE	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Albania	AL	RIPE	0	0	0	0	0	0	1	0	0	0	0	1	2	0
Algeria	DZ	AFRINIC	0	0	0	0	1	0	1	0	0	1	0	7	2	0
American Samoa	AS	APNIC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Andorra	AD	RIPE	1	0	0	1	1	0	1	0	0	1	0	0	1	0
Angola	AO	AFRINIC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anguilla	AI	ARIN	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Antarctica	AQ	ARIN	1	1	1	0	0	0	1	0	1	0	0	0	0	0
Antigua and Barbuda	AG	ARIN	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Argentina	AR	LACNIC	1	1	1	0	1	0	0	2	1	0	0	1	1	2

Figura 6. BASE-2. Fonte: http://ietf.protocolos.net.br/meetings/ietf.php

A terceira base, *BASE-3* foi obtida, de forma acumulativa, a partir da *BASE-2* tendo como foco principal, a região ("RIR"), conforme a Figura 7.

	2008		2009			2010			2011			2012		
RIR	72	73	74	75	76	77	78	79	80	81	82	83	84	85
AFRINIC	9	5	31	34	6	30	20	24	19	23	7	45	52	52
APNIC	335	172	258	288	629	306	350	646	291	300	424	318	325	288
ARIN	494	517	730	441	351	723	434	390	499	630	388	574	670	720
LACNIC	10	10	12	14	12	12	16	22	15	12	13	12	24	24
RIPE	470	255	305	454	249	282	490	261	498	284	224	587	288	270
Totals	1318	959	1336	1231	1247	1353	1310	1343	1322	1249	1056	1536	1359	1354
	2277		3814			4006			3627			4249		

Figura 7. BASE-3. Fonte: http://ietf.protocolos.net.br/meetings/ietf.php

3.2. As atividades de pré-processamento sobre as bases de dados originais

Dois tipos de pré-processamento foram executados com o objetivo de procurar por padrões e por indicações ou hipóteses de quais as tarefas de mineração trariam resultados adequados para uma avaliação eficaz das bases originais: pré-processamento externo e pré-processamento interno. O primeiro se refere às atividades executadas antes do *RapidMiner* e o segundo envolve diretamente o uso *RapidMiner* como ferramenta de pré-processamento.

3.2.1. Pré-processamento externo

Um programa em PHP capturou o atributo *ISO_3166 Code* da base original (Figura 1, página 1), identificando em uma tabela MySQL qual região ele pertencia, acumulando por encontro. Isto ocorreu para os encontros de número 72 a 85. A partir desta base MySQL, foram obtidas as bases *BASE-2*, e *BASE-3*. Estas duas bases foram copiadas para o Excel, em abas diferentes, juntamente com a *BASE-1*, referente a reunião IETF 89, única considerada neste estudo.

3.2.2. Pré-processamento interno

O processamento interno (no *RapidMiner*) ocorreu em várias etapas, repetitivamente. Decidiu-se, preliminarmente, pela avaliação conjunta das três bases, conforme mostram a Figura 8.

A Figura 8 exibe o resultado no *RapidMiner*, como se pode ver no cabeçalho e no detalhamento dos três resultados em *Process*, dentro da *Perspectiva de Resultados*. Esta facilidade do *RapidMiner* é muito conveniente e pode ir um pouco além. Por exemplo, se o trabalho estiver concentrado na *BASE-1*, basta retirar as conexões finais das outras bases para que somente os resultados da *BASE-1* sejam mostrados.

Alguns atributos não foram importados, a saber: na *BASE-1*, o atributo *Country* e na *BASE-2*, o atributo *País*, pois ambos são equivalentes ao atributo *ccTLD*. Após a importação das bases, dois indicadores foram usados para uma análise preliminar: o *Meta Data View* (*MDV*) e o *Plot View* (*PV*). No *PV*, entre as diversas opções disponíveis e avaliadas, preferiu-se exibir o diagrama de espalhamento.

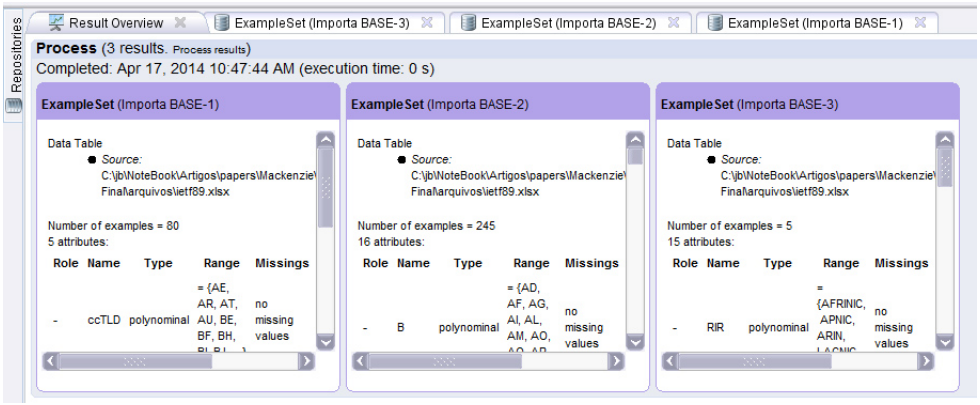
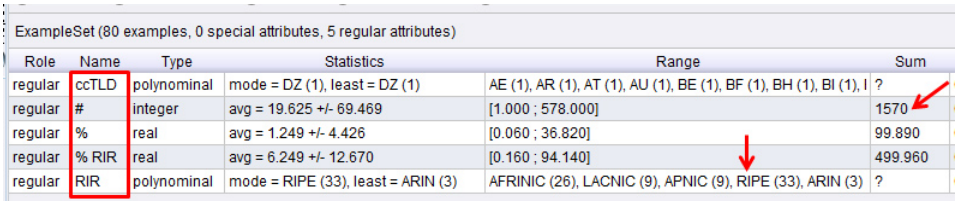


Figura 8. As três bases de dados originais no RapidMinder

BASE-1 O resultado da operação *Importa BASE-1* indicou o *MDV* exibido na Figura 9, e o diagrama de espalhamento mostrado na Figura 10, com algumas observações itemizadas, na sequência.



ExampleSet (80 examples, 0 special attributes, 5 regular attributes)

Role	Name	Type	Statistics	Range	Sum
regular	ccTLD	polynominal	mode = DZ (1), least = DZ (1)	AE (1), AR (1), AT (1), AU (1), BE (1), BF (1), BH (1), BI (1), I	?
regular	#	integer	avg = 19.625 +/- 69.469	[1.000 ; 578.000]	1570
regular	%	real	avg = 1.249 +/- 4.426	[0.060 ; 36.820]	99.890
regular	% RIR	real	avg = 6.249 +/- 12.670	[0.160 ; 94.140]	499.960
regular	RIR	polynominal	mode = RIPE (33), least = ARIN (3)	AFRINIC (26), LACNIC (9), APNIC (9), RIPE (33), ARIN (3)	?

Figura 9. "Meta Data View"da BASE-1

Em relação ao MDV mostrado na Figura 9 pode-se comentar:

MDV1.a. O atributo *Country* foi retirado na fase de importação. , já que o atributo *ccTLD* pode representá-lo.
MDV1.b. Na coluna *SUM* o número 1570 (inscritos) está compatível com a base original, na página do IETF.
MDV1.c. Vale notar que, a região do *ARIN* (entre parênteses, na coluna *Range*, do atributo *RIR*, trinta e três países enviaram representantes. No *MDV* da *BASE-3* (Figura 13, página 9), referente a um único encontro, somente três países do *ARIN* estiveram presentes. O encontro 89 do IETF aconteceu em Londres..
MDV1.d. Os nomes dos atributos não estão visivelmente adequados e devem ser alterados com nomes mais representativos.

Em relação ao diagrama de espalhamento mostrado na Figura 10, comenta-se:

PV1.a. Este diagrama é, praticamente, um agrupamento de participantes por RIR. Isto foi conclusivo em uma experiência, no *RapidMiner*, com o operador *Clustering* sobre este conjunto de dados.
PV1.b. Os agrupamentos apontados nos círculos, a primeira vista podem parecer *discrepantes*. Considerando a fonte de dados, que mede as inscrições no encontro 89 do IETF, não se pode concluir desta forma. É apenas um número de inscrições bem acima da média, mas válido.
PV1.c. Atento ao agrupamento formado pelo RIPE observa-se que o local da reunião influencia na participação. Os dois pontos, presumivelmente discrepantes referem-se a encontros ocorridos em sua região de influência.
PV1.d. A observação do item anterior aplica-se, também , ao APNIC.

BASE-2 O resultado de importação da *BASE-2*, resultou no *MDV* mostrado na Figura 11 e, ao diagrama de espalhamento, na Figura 12, com algumas informações abordadas na sequência.

Em relação ao MDV mostrado na Figura 11:

MDV2.a. O atributo *País* foi retirado, sob o argumento utilizado no item (MDV1.a.) da *BASE-1*.
MDV2.b. A coluna "Sum"representa a participação total em cada encontro do IETF.
MDV2.c. O atributo RIR contrasta com a observação do item MDV1.c., na coluna *Range*.Os eventos de 72 a 85, vê-se que 28 países da região do ARIN compareceram ao conjunto dos 14 eventos. ARIN é a região da

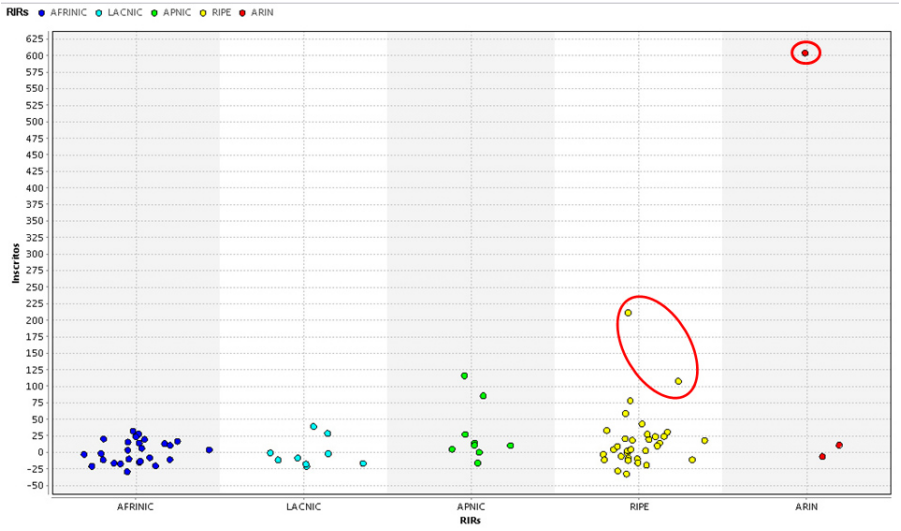


Figura 10. Diagrama de espalhamento da *BASE-1*

ExampleSet (243 examples, 0 special attributes, 16 regular attributes)

Role	Name	Type	Statistics	Range	Sum
regular	ccTLD	polynominal	mode = AF (1), least = AF (1)	AD (1), AE (1), AF (1), AG (1), AI (1), AL (1), AM (1), AN (1), AO (	?
regular	RIR	polynominal	mode = RIPE (76), least = ARIN (28)	APNIC (54), RIPE (76), AFRINIC (56), ARIN (28), LACNIC (29)	?
regular	72	integer	avg = 5.424 +/- 31.783	[0.000 ; 453.000]	1318
regular	73	integer	avg = 3.947 +/- 32.325	[0.000 ; 489.000]	959
regular	74	integer	avg = 5.498 +/- 45.331	[0.000 ; 689.000]	1336
regular	75	integer	avg = 5.066 +/- 30.210	[0.000 ; 417.000]	1231
regular	76	integer	avg = 5.132 +/- 35.604	[0.000 ; 430.000]	1247
regular	77	integer	avg = 5.568 +/- 45.220	[0.000 ; 681.000]	1353
regular	78	integer	avg = 5.391 +/- 30.072	[0.000 ; 400.000]	1310
regular	79	integer	avg = 5.527 +/- 37.115	[0.000 ; 437.000]	1343
regular	80	integer	avg = 5.440 +/- 32.750	[0.000 ; 462.000]	1322
regular	81	integer	avg = 5.140 +/- 37.702	[0.000 ; 554.000]	1249
regular	82	integer	avg = 4.346 +/- 27.103	[0.000 ; 367.000]	1056
regular	83	integer	avg = 6.321 +/- 38.496	[0.000 ; 535.000]	1536
regular	84	integer	avg = 5.593 +/- 41.093	[0.000 ; 611.000]	1359
regular	85	integer	avg = 5.572 +/- 44.984	[0.000 ; 679.000]	1354

Figura 11. "Meta Data View"da *BASE-2*

América do Norte, (Figura 4, página 5). Uma hipótese é de que boa parte dos eventos tenha ocorrido nesta região (Figura 3, página 3) facilitando o trânsito dos interessados em participar.

MDV2.d. A coluna *Name* possui nomes começando com números. O *RapidMiner* tem restrições em relação a este tipo de nome, caso sejam usados em expressões aritméticas. Portanto, seria adequado renomeá-los, incluindo os nomes *ccTLD* e *RIR*, apropriadamente. Na troca de nomes, a indicação dos eventos deve incluir o nome da localidade permitindo uma visão mais direta e compreensível. Localidade é uma propriedade importante, dos dados originais e, fundamental, em análises subsequentes.

Em relação ao diagrama de espalhamento mostrado na Figura 12 diz-se:

PV1.a. Os pontos circulados representam, na sua grande maioria, inscrições em eventos ocorridos na respectiva região e, portanto, como já dito, não são *outliers*.

PV1.b. A região do AFRINIC possui mais inscritos do que a região do LACNIC. Muitos eventos do IETF realizaram-se na Europa, local de trânsito mais fácil para a região do LACNIC. Nestes eventos, a presença do AFRINIC é maior, como se pode observar pelos agrupamentos marcados.

BASE-3 A importação da *BASE-3* resultou no "MDV"mostrado na Figura 13, com algumas indicações, em vermelho.

Pode-se comentar:

MDV3.a. A coluna "Sum"está compatível com a mesma coluna, da BASE-2

MDV3.b. Os nomes dos atributos foram alterados, seguindo a sugestão do item MDV2.d.

Em relação ao diagrama de espalhamento mostrado na Figura 14:

Figura 12. Diagrama de espalhamento da *BASE-2*

ExampleSet (5 examples, 0 special attributes, 15 regular attributes)

Role	Name	Type	Statistics	Range	Sum
regular	RIRs	polynominal	mode = AFRINIC (1), least = AFRINIC (1)	AFRINIC (1), APNIC (1	?
regular	a72_Dublin	integer	avg = 263.600 +/- 239.748	[9.000 ; 494.000]	1318
regular	a73_Minneapolis	integer	avg = 191.800 +/- 211.002	[5.000 ; 517.000]	959
regular	a74_SanFrancisco	integer	avg = 267.200 +/- 290.093	[12.000 ; 730.000]	1336
regular	a75_Stockholm	integer	avg = 246.200 +/- 213.202	[14.000 ; 454.000]	1231
regular	a76_Hiroshima	integer	avg = 249.400 +/- 259.818	[6.000 ; 629.000]	1247
regular	a77_Anaheim	integer	avg = 270.600 +/- 287.581	[12.000 ; 723.000]	1353
regular	a78_Maastricht	integer	avg = 262 +/- 228.250	[16.000 ; 490.000]	1310
regular	a79_Beijing	integer	avg = 268.600 +/- 263.566	[22.000 ; 646.000]	1343
regular	a80_Prague	integer	avg = 264.400 +/- 241.213	[15.000 ; 499.000]	1322
regular	a81_Quebec	integer	avg = 249.800 +/- 253.095	[12.000 ; 630.000]	1249
regular	a82_Taipei	integer	avg = 211.200 +/- 198.549	[7.000 ; 424.000]	1056
regular	a83_Paris	integer	avg = 307.200 +/- 276.351	[12.000 ; 587.000]	1536
regular	a84_Vancouver	integer	avg = 271.800 +/- 260.467	[24.000 ; 670.000]	1359
regular	a85_Atlanta	integer	avg = 270.800 +/- 278.774	[24.000 ; 720.000]	1354

Figura 13. "Meta Data View"da BASE-3

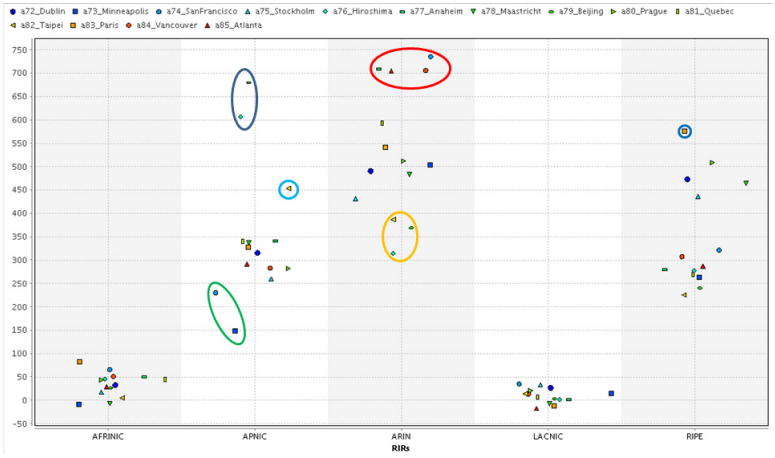


Figura 14. Diagrama de espalhamento da BASE-3

PV1.a. Apesar da pouca participação, a proximidade com os locais de realização dos eventos influencia o maior número de participantes do AFRINIC em relação ao LACNIC.

PV1.b. ARIN, RIPE e APNIC são visivelmente afetados pelo local de realização dos eventos. Entretanto, o ARIN é mais frequente, independente da localidade do evento.

PV1.c. Os diagramas de espalhamento, de todas as bases indicam que, realmente, o local é preponderante nas inscrições, embora a região do ARIN mantenha-as em agrupamentos equilibrados .

3.2.3. Considerações finais sobre a etapa de pré-processamento

O mecanismo exaustivo, de iteração e interação descrito por [Fayyad et al. 1996] ocorrerá, nas etapas de análise dos dados e refinamentos necessários em parâmetros envolvendo os algoritmos de mineração. Embora seja pouco provável a necessidade de pré-processamento externo, a execução repetitiva na ferramenta será evidente, quando da otimização dos resultados, ajustando-os a melhores condições nas análises finais.

4. Resultados

Na análise dos resultados, as atividades de pré-processamento foram executadas no *RapidMiner*, repetidamente. Isto ocorreu sobre todas as bases e foi considerado experimental, de tal forma que somente as atividades relacionadas com a *BASE-1* foram reproduzidas no presente trabalho.

4.1. BASE-1

No contexto do evento 89 pode-se imaginar uma forma de classificar os países, nas respectivas regiões, pelo número de inscrições. Considerando que 80 exemplares é um número razoável, a classificação através de uma árvore poderia ser a escolha preferida, principalmente pelo seu resultado intuitivo. Resta saber se é uma tarefa adequada aos objetivos desejados.

4.1.1. Árvore de decisão

Os mineradores sempre procuram árvores de decisão menos complexas, pois elas são mais compreensíveis para o uso humano. No caso da *BASE-1*, esta não é uma preocupação relevante, diante do número de exemplares (80). Mas, devem ser levadas em conta, as recomendações de [Rokach and Maimon 2005], de que a complexidade de uma árvore de decisão possui as seguintes métricas:

CA.a. O número total de nodos.
CA.b. O número total de folhas,
CA.c. A profundidade da árvore.
CA.d. O número de atributos usados.

No *RapidMiner*, os itens CA.a., CA.b. e CA.c. são controlados, respectivamente, pelos parâmetros *minimal size for split*, *minimal leaf size* e *maximal depth*, do operador *Decision Tree*. O item CA.d. é controlado manualmente, pela escolha dos atributos no momento da importação dos dados de treinamento ou logo após, através de alguns operadores do *RapidMiner*, disponíveis para este tipo de transformação. Logo após a importação dos dados, foi usado o operador *Select Attributes* e na sequência, o operador *Set Role* que transformou o atributo *RIRs* em rótulo (ou classe). Diversas experiências foram feitas em relação ao item CA.d. e como se esperava, a melhor *acurácia* foi aquela com os atributos dos dados de treinamento: (*Inscritos*, *Total_%* e *RIR_%*). Um outro parâmetro utilizado na tarefa de classificação usando árvore de indução é a *poda* (*prune*), usada para remover possíveis anormalidades nos dados de treinamento. Foi desabilitada esta opção, pois o diagrama de espalhamento da *BASE-1* mostrou que não se pode considerar como discrepância,

algumas instâncias fora dos agrupamentos mais consistentes (Figura 10, página 8), e comentado no item PV1.b.. O processo operacional utilizando o *RapidMiner* e descrito, complementarmente a seguir está representado na Figura 15.

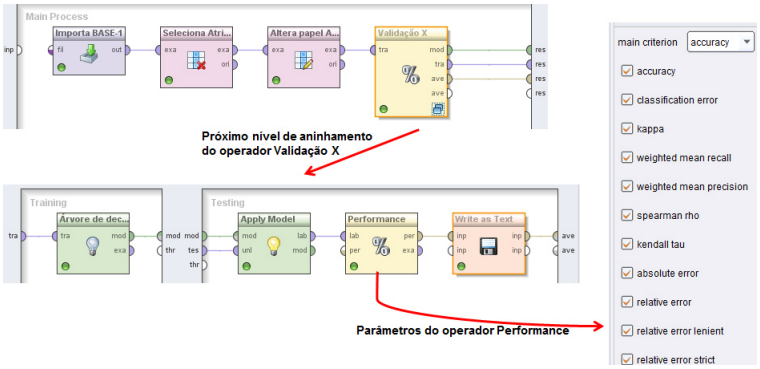


Figura 15. O processo de execução da árvore de decisão no *RapidMiner*.

Se na sequência, isto é, após o operador *Set Role* fosse inserido o operador *Decision Tree* o resultado da árvore é o que se vê na Figura 16. Esta solução, imediata, não traria nenhuma medida de eficiência junto com a árvore resultante e portanto, o minerador ficaria sem informações sobre o resultado desta classificação. Para isto é necessário estabelecer o *conjunto de testes*. Isto é feito pelo operador de *Validação*. A escolha foi pelo operador *X Validation*, do *RapidMiner*, que treina o modelo preditivo (a árvore) sobre o conjunto de treinamento e testa-o sobre o conjunto de teste (maior do que nos outros operadores de validação), oportunidade em que as medidas de eficiência são executadas. Este operador possui um parâmetro, o *number of validations* sobre o qual será falado, mais a frente.

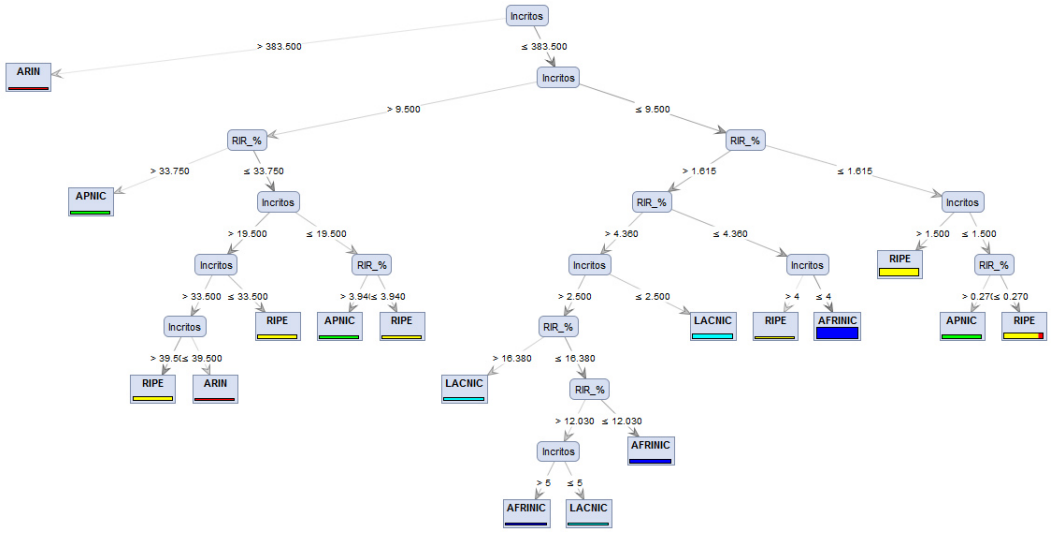
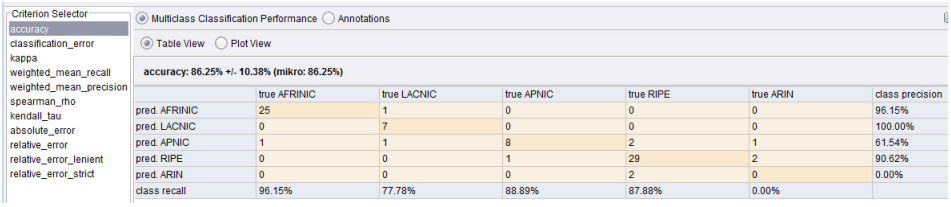


Figura 16. Árvore de decisão da BASE-1.

O operador *X Validation* é aninhado, o que significa que a tarefa *Decision Tree* deverá ocorrer no seu próximo nível de aninhamento, o *Training*. Neste mesmo nível existe o ambiente de *Testing*, no qual será usado o operador *Apply Model*, com efeitos

de predição aplica o modelo sobre um conjunto exemplo. O resultado desta aplicação será encaminhada para o operador *Performance (Classification)*, que oferece um conjunto de parâmetros de avaliações bastante completo. Complementarmente, foi usado o operador *Write as Text* que grava em um arquivo texto, o resutado do operador *Performance (Classification)*, para cada uma das avaliações definidas no parâmetro *number of validations* do operador *X Validation*. Os parâmetros resultantes e a *matriz de confusão* estão exibidos na Figura 17.



	true AFRINIC	true LACNIC	true APNIC	true RIPE	true ARIN	class precision
pred. AFRINIC	25	1	0	0	0	96.15%
pred. LACNIC	0	7	0	0	0	100.00%
pred. APNIC	1	1	8	2	1	61.54%
pred. RIPE	0	0	1	29	2	90.62%
pred. ARIN	0	0	0	2	0	0.00%
class recall	96.15%	77.78%	88.89%	87.88%	0.00%	

Figura 17. Matriz de Confusão da árvore da Figura 16.

O processo de predição (ou classificação) da árvore de decisão, depende de dois parâmetros: *Inscritos* e *RIR_%*. Por exemplo, a Nigéria, com 5 inscritos e 10,42 na frequência de RIR (*RIR_%*). A partir da raiz, a sequência da Figura 18 mostra que a ávore classifica a Nigéria como situada na região do AFRINIC, o que era de se esperar.

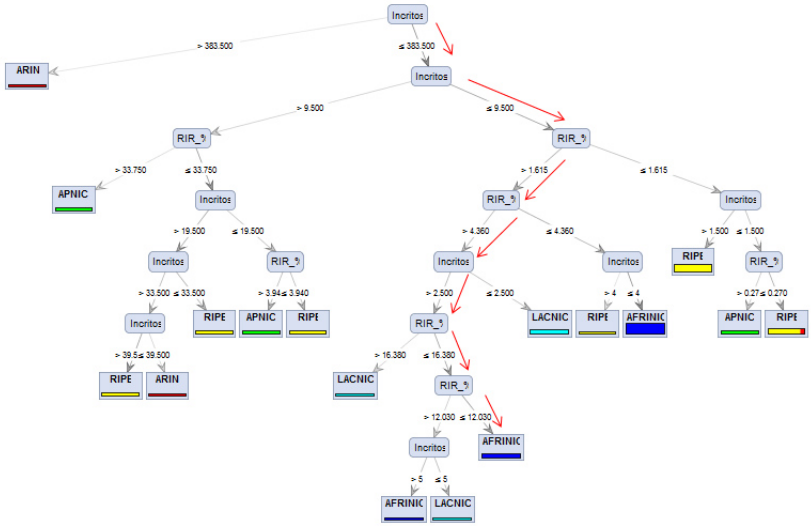


Figura 18. Classificando a Nigéria, com cinco (5) inscritos.

A questão mais importante, após esta verificação do exemplo da Nigéria é responder a seguinte pergunta: a resposta da classificação é adequada para qualquer exemplar da *BASE-1*, ou a árvore obtida classifica inteqralmente a *BASE-1*, sem erro?

A matriz de confusão e os parâmetros podem responder a esta pergunta. Porém, a árvore da Figura 16 já dá algumas indicações nos *nós terminais*. O *RapidMiner* desenha a árvore com cores indicativas em seus nós terminais. Todos eles, exceto o nó terminal mais à direita, correspondente a RIPE, exibem somente uma cor. No formato texto da árvore ou ao colocar o mouse sobre este nó terminal, ele mostra

o número de instâncias de exemplares chegam a este nó, como: *AFRINIC=0, LACNIC=0, APNIC=0, RIPE=7, ARIN=1*. Observa-se que há uma instância do ARIN chegando a este nó e, para a qual, a árvore deverá classificar com erro (isto é, não conseguirá decidir). O minerador deve ponderar a respeito deste fato e, se considera ou não a árvore como adequada aos objetivos da classificação.

Na matriz de confusão, os valores da diagonal representam os acertos e os demais valores, os erros. Assim, eis a interpretação da matriz de confusão resultante em relação ao modelo da árvore de decisão mostrada na Figura 16, em uma leitura horizontal: (a) O modelo prediz corretamente, 25 vezes para AFRINIC e incorretamente, 1 vez, favorável ao LACNIC. A predição estará correta, portanto, em 96,15% das vezes, (b) O modelo prediz corretamente, 7 vezes para o LACNIC e não há nenhuma predição incorreta. Garante, portanto, 100% de acertos, (c) O modelo prediz corretamente, 8 vezes para o APNIC e incorre em 5 erros de classificação. Consegue garantir 61,54% de acertos para o APNIC, (d) Para o RIPE, o modelo acerta 29 vezes e erra 3 vezes. Isto equivale a 90.62% de acertos e, (e) Para o ARIN, o modelo não consegue predizer nenhum acerto e pode levar duas vezes ao erro. Portanto: 0% de acerto.

Para confirmar a imprecisão do modelo de árvore resultante, bastaria verificar o exemplar *Jamaica*, cujo resultado seria RIPE, e não ARIN, como deveria.

Embora a matriz de confusão não seja de interpretação fácil o minerador examinou em detalhes o arquivo texto produzido pelo operador *Write as Text*. Este arquivo registra as matrizes de confusão para cada um das validações definidas no parâmetro correspondente do operador *Validação X*, cujo valor foi 10. A matriz de confusão da Figura 17 é, na realidade, a soma de todas as matrizes intermediárias gravada no arquivo referido. Assim, a única conclusão possível para o minerador é de que da *BASE-1* ou, do conjunto de treinamento, da forma original é um domínio para o qual uma tarefa de classificação não poderá ser usado com resultados ótimos, ou melhor, que faça a predição sem nenhum erro. A resposta à pergunta sobre a precisão da árvore é nâo!

Mas, o trabalho do minerador é entregar uma árvore de decisão eficaz para a *BASE-1*. Há uma solução, com base no argumento referido no item CA.d., proposto por [Rokach and Maimon 2005]. O minerador imagina que se adicionar mais um atributo auxiliar daria força ao modelo, corrigindo a ineficiência demonstrada pela matriz de confusão acima. Na Figura 19 este operador é o *Generate Attributes*.

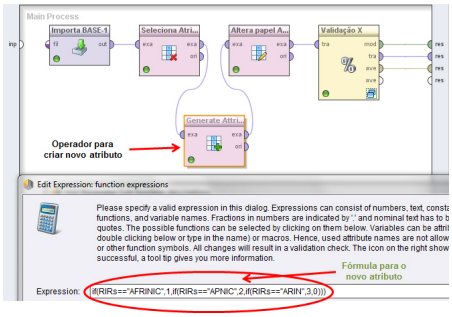


Figura 19. Novo atributo e respectiva fórmula de criação.

A fórmula para gerar o novo atributo foi conseguida após várias experiências. O resultado foi a árvore de decisão da Figura 20 com a respectiva tabela de confusão, na Figura 21.

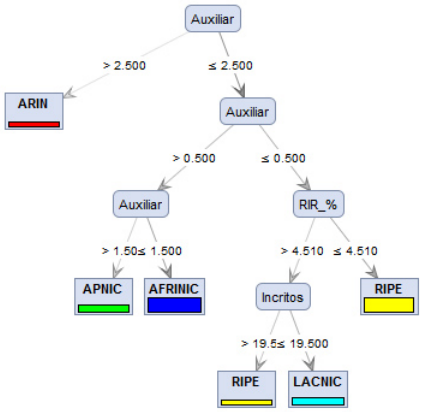


Figura 20. Árvore de classificação para a *BASE-1*, após a inclusão do atributo *Auxiliar*.

accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

	true AFRINIC	true LACNIC	true APNIC	true RIPE	true ARIN	class precision
pred. AFRINIC	26	0	0	0	0	100.00%
pred. LACNIC	0	9	0	0	0	100.00%
pred. APNIC	0	0	9	0	0	100.00%
pred. RIPE	0	0	0	33	0	100.00%
pred. ARIN	0	0	0	0	3	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	

Figura 21. Matriz de Confusão da árvore da Figura 20.

4.1.2. Análise da matriz de confusão e demais indicadores do modelo

Para efeitos desta análise será usada a matriz de confusão mostrada na Figura 17 na página 12, que adicionalmente exibe a listas dos índices calculados, à esquerda. Da matriz de confusão conclui-se: (a) O modelo faz 69 predições corretas $(25+7+8+29+0)$, (b) O modelo faz 11 predições incorretas $(1+1+1+1+2+2+1 = 80-69)$, (c) A *BASE-1* possui 80 exemplares, (d) A taxa de erro é $11/80 = 0,14$ e, (e) A acurácia é de $69/80 = 0,8625$, conforme se vê na matriz.

Os indicadores[14] mais significativos são: *accuracy*, referente à percentagem de predições corretas e o *classification_error*, referente à percentagem de predições incorretas. Seus resultados na primeira matriz de confusão são, respectivamente: (accuracy: 86.25% +/- 10.38%) e (classification_error: 13.75% +/- 10.38%). O valor após os sinais $+/-$ reperesenta o desvio padrao. Todos os indicadores representam a média aritmética entre os resultados intermediários, registrados pelo operador *Write as Text*, no arquivo definido.

5. Conclusão

Pelo que se obervou, o local de realização da reunião estimula a participação de membros do AFRINIC, APNIC, ARIN e RIPE. Nada se pode concluir em relação ao

[14] `http://www.rapid-i.com`

LACNIC, pois não houve nenhum evento abaixo da linha do Equador(`http://www.ietf.org/meeting/upcoming.html`), exceto na Austrália, que está bem afastada desta região. Isto poderá ser confirmado na reunião que se realizará em Buenos Aires, em 2016, provavelmente.

No diagrama de espalhamento (Figura 14), os três eventos de 2013 e o primeiro evento de 2014, não estão contemplados. Tal inclusão aumentaria o conjunto de treinamento referente a *BASE-2* afetanto, para melhor, o resultado das análises, incluindo a perspectiva de novas tarefas de mineração. Dito de forma mais precisa, o *RapidMiner* pode auxiliar a alteração na *BASE-2*, na inclusão dos dados relativos aos eventos 86 a 89 e, ato contínuo reconstruir a *BASE-3* e atualizar o atributo *Inscritos* e respectivos percentuais referentes aos outros atributos, da *BASE-1*. Neste raciocínio, a *BASE-1* e *BASE-3*, poderão ser eliminadas do processo de mineração, já que a *BASE-2* é o resultados dos dados capturados, na origem. É uma sequência interessante, pois a cada novo evento seriam incluidos os novos dados referente às inscrições, não afetando as escolhas das tarefas de mineração mas, tão somente, o pré-processamento. Outra atividade de avaliação dos resultados é considerar como dados de treinamento, outras bases equivalentes à *BASE-3*, para serem usadas como conjuntos de testes em resultados de classificação. A continuação dos estudos relacionados e propostos para o futuro terão continuidade, por exigência de um esforço mais abrangente para ampliar a aquisição e guarda de conhecimento na Infraestrutura da Internet, em pesquisa proposta por [Braga and Omar 2014].

O *RapidMiner*, como ferramenta de mineração é bastante flexível, principalmente se forem adicionadas, as facilidades do sistema R e o complemento do Weka, entre outros. Tais recursos estimulam a atratividade da aplicação de Mineração de Dados em bases de dados da Infraestrutura da Internet.

Referências

Braga, J. and Omar, N. (2014). Semantic repository in internet infrastructure knowledge domain: Methodology. In *CSBC 2014 - IWPIETF LAC*, http://iwpietf.lacnog.org/. Disponível em `https://jems.sbc.org.br/PS.cgi/128952.1.pdf?m=128952\&fi=1\&f=128952_1.pdf`.

Chisholm, A. (2013). *Explore, understand, and prepare real data using RapidMiner's practical tips and tricks*. Packt Publishing Ltd., Birmingham, UK, 1 edition.

Fayyad, U., Piatetsky-shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17:37–54.

Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers: A survey. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, Vol. 35, No. 4.

Tan, P., Steinbach, M., and Kumar, V. (2009). *Introdução ao Data Mining: Mineração de Dados*. Editora Ciência Moderna, Rio de Janeiro.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques*. Elsevier, USA, 3rd edition.