# Construção de Conjunto de Classificadores Baseado na Diversidade do Espaço de Características e Algoritmos de Aprendizagem para Detecção de Spam

**Gabriel P. Lutz**[1], **Lucas Ost**[1], **Márcia Henke**[1]

[1]Colégio Técnico Industrial de Santa Maria - Universidade Federal de Santa Maria (UFSM)
Santa Maria – RS – Brazil

{gabiplutz,lucasost,henke}@redes.ufsm.br

***Abstract.*** *Research in the class machines area focus their efforts on diversity for the construction ensemble classifiers. The concept of diversity is related to the resources used to develop ensemble classifiers. It is demonstrated that diversity over learning algorithms performs better than feature manipulation capabilities. Showing considerable reduction of false positives in problem spam classification, in addition to the other metrics addressed as precision, accuracy, measure-f1 and recall.*

***Resumo.*** *As pesquisas na área de aprendizagem de máquina estão focando seus esforços na diversidade para construção de conjunto de classificadores. O conceito de diversidade está relacionado aos recursos usados para formar um conjunto de classificadores. Este trabalho apresenta experimentos considerando manipulação de características/instâncias e algoritmos de aprendizagem. Demonstra-se que a diversidade considerando algoritmos de aprendizagem tem um desempenho superior aos recursos de manipulação de características. Apresentando considerável redução de falsos positivos na classificação binária de spam, além das demais métricas abordadas como precisão, acurácia, medida-f1 e recall.*

## 1. Introdução

O contínuo aumento no número de dispositivos conectados à Internet tem estimulado criminosos cibernéticos na sofisticação dos *softwares* e ferramentas maliciosas para conduzir ataques e intrusões de computadores usando e-mails não solicitados, denominado *spam*, como maior vetor de disseminação. Cerca de 55,4% de todos os *e-mails* que circulam diariamente são identificados como *spam* [Symantec 2018].

A partir deste contexto grupos de pesquisas buscam aperfeiçoar os métodos de detecção de *spam*. Construção de conjunto de classificadores aplicando o conceito diversidade tem sido o foco de muitas pesquisas na área de segurança e aprendizagem de máquina. A aplicação do conceito de diversidade está relacionado na forma que se pode gerar um conjunto de classificadores, denominados de conjunto de classificadores homogêneos e heterogêneos. Conjuntos de classificadores heterogêneos, empregam diferentes algoritmos de aprendizagem, treinados sobre um mesmo conjunto de características, o que torna o aprendizado diversificado e a maneira com que cada classificador aprende o problema. [Easwaramoorthy et al. 2016] apresenta um modelo de classificação combinando três classificadores base: *Naive Bayes* (NB), *Multilayer Perceptron* (MLP)

e *Decision Tree*(DT). Seguindo a mesma linha de pesquisa [Ibrahim et al. 2017] empregam outros três classificadores *Naive Bayes*, *Support Vector Machine* (SVM) e *Logistic Regression* (LR), onde é realizada uma votação ponderada para definir o rótulo de classe final.

Classificadores homogêneos, empregam estratégias de redução da dimensionalidade de características, onde cada amostra do conjunto de dados de treino é diferente, dando a cada classificador treinado, foco e perspectiva sutilmente diferentes sobre o problema. [Yin et al. 2014] optam pela construção de um conjunto de classificadores considerando algoritmos que lidam com diversidade, como Algoritmos Genético (GA), utilizando métodos convencionais para geração de conjunto de classificadores como: *Bagging*, *AdaBoost*, entre outros. [Bhat et al. 2014] aplicando *Bagging*, *Boosting* e *Stacking*, que faz a combinação de diversos classificadores base (classificador base é um classificador individual que irá compor o conjunto de classificadores) através de voto majoritário. [Díez-Pastor et al. 2015] aplica, também, técnicas convencionais como: *Bagging*, *Boosting*, *Random Balance* e SMOTE, baseadas em manipulação de instâncias, e combinam técnicas de *Rotation Forest*, *Random Feature Weights* e *Disturbing Neighbours*, baseadas em manipulação de características.

Este trabalho apresenta a duas séries de experimentos baseado na construção de conjunto de classificadores homogêneos e heterogêneos. Este trabalho emprega algoritmos de aprendizagem e manipulação de características/instâncias para construir um conjunto de classificadores. Para tanto os classificadores base empregados são três: *Naive Bayes* (NB), *Support Vector Machine* (SVM) e *Logistic Regression* (LR).

Diferentemente das abordagens apresentadas este trabalho tem como objetivo principal determinar um limiar para um novo retreino, sem considerar a taxa de erro do modelo de classificação. Utilizando o cálculo do cosseno para investigar a similaridade das características que compõem o conjunto de treino em relação as características que compõem o conjunto de teste. Acredita-se que quanto menor a similaridade entre características maior a taxa de erro. Logo o retreino ficaria condicionado a taxa de similaridade e não a taxa de erro.

## 2. Protocolo Experimental

Esta seção descreve os experimentos realizados para avaliar as abordagens investigadas empregando a ferramenta *WEKA*.

### 2.1. Base de dados

Os experimentos deste trabalho foram realizados utilizando a base de dados: ECUE (*Email Classification Using Examples*) - base pública disponibilizada na web em http://www.dit.ie/computing/staff/sjdelany/datasets/. A base de dados é composta por uma coleção de amostras *spam* e *e-mails* legítimos recebidos no período de 2002 a 2004. A distribuição dos dados da base de treino é composta por 1000 amostras (500 *spams*, 500 legítimas), e da base de teste é composta por 1289, sendo 938 amostras *spam* e 351 amostras legítimas.

### 2.2. Classificadores Base e Medidas de Desempenho

Para gerar o conjunto de classificadores foram empregados três classificadores base: *Naive Bayes* (NB), *Support Vector Machine* (SVM) e *Logistic Regression* (LR)

[Ibrahim et al. 2017]. As métricas empregadas se fundamentam na matriz de confusão. A matriz de confusão é composta pelas seguintes taxas de classificação: falso positivo (*False positive* - FP), classificação de *e-mails* legítimos como *spam*; falso negativo (*False negative* - FN), mensagens com conteúdo spam classificadas como não-*spam*. Verdadeiro positivo (*True positive* – TP) *spam* classificado corretamente como *spam* e verdadeiro negativo (*True negative* – TN). As demais métricas aplicadas se utilizam das informações apresentadas na matriz de confusão: (i) acurácia (*Accuracy*); (ii) precisão (*Precision*); sensibilidade (*Recall*) e Medida-F1(*F1-measure*) [Ibrahim et al. 2017].

### 2.3. Experimentos

Para realização dos experimentos foi empregada a ferramenta WEKA. Os experimentos empregam duas abordagens para construção de conjunto de classificadores agregando três classificadores base (SVM, NB e LR). A partir deste contexto são geradas duas séries de experimentos:

**Baseada na manipulação do espaço de características (Série 1):** na primeira série cada amostra do conjunto de dados de treinamento é diferente, dando a cada classificador treinado, foco e perspectiva sutilmente diferentes sobre o problema. A estratégia de redução de dimensionalidade de características aplicada foi RSM (*Random Subspace Method*) [Ibrahim et al. 2017].

**Baseada em diferentes algoritmos de aprendizagem (Série 2):** na segunda série cada amostra do conjunto de treino é o mesmo, o que torna o aprendizado diversificado é a maneira com que cada classificador aprende o problema. A estratégia de redução de dimensionalidade de características aplicada foi *Info Gain* (IG) [Ibrahim et al. 2017].

Os parâmetros de configuração para os classificadores NB e LR foram mantidos os padrões fornecidas pela ferramenta *WEKA*. Para o classificador SVM foi realizado uma bateria de experimentos sobre a base de treino para definir os valores com melhor desempenho para o fator de penalização e o y do kernel, sendo o fator de penalização alterado para C=5 e o kernel utilizado foi o Radial Basis Function (RBF) com y=0.01. Tanto a série 1 quanto a série 2 unem a decisão de cada classificador a partir do voto majoritário. A quantidade de classificadores combinados em ambas as séries são de três classificadores base.

O conjunto de classificadores gerado a partir da série 1, construiu três conjuntos de classificadores homogêneos. A série 2 gerou um conjunto de classificadores heterogêneos.

## 3. Resultados e Trabalhos Futuros

A estabilidade apresentada pelo conjunto de classificadores gerados na Série 2 se reflete nas métricas apresentadas na Figura 1. Observa-se que os valores de desempenho da Série 1, Figura 1 (a), está entre 98,10% a 94,88%, enquanto a Série 2 apresenta estabilidade com 98,30% para todas métricas. Esse comportamento pode estar relacionado com a quantidade de FP e FN em que a Série 2 apresenta uma considerável redução, Figura 1 (c). Para finalizar a Figura 1 (b) apresenta a Série 2 com a acurácia melhor que a Série 1, com percentual de 98,52%. Importante salientar que os valores reduzem na quantidade de FP, alcançados pela Série 2, em até 0,85% em relação Série 1, respectivamente: 1,71%; 3,99% e 1,42%. Os resultados alcançados ajudaram a direcionar os próximos passos com

conjunto de classificadores heterogêneos. Segue-se no desenvolvimento de um método dinâmico para substituição de um dos classificadores base na presença da redução de seu desempenho.

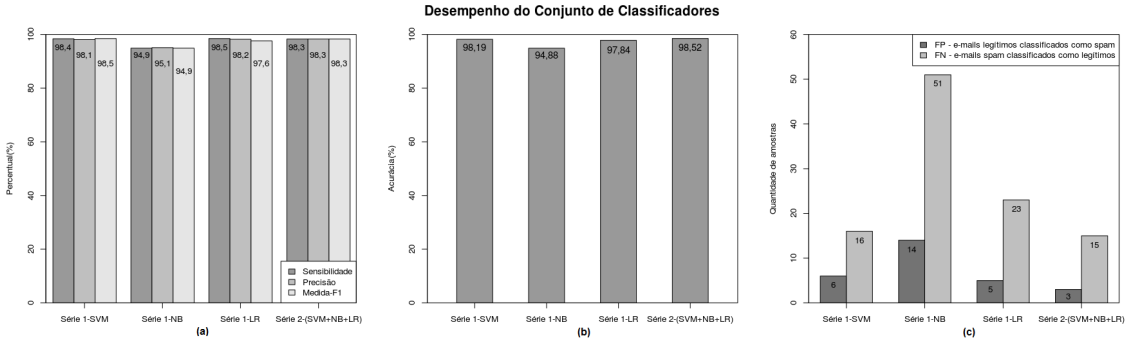


**Figura 1. Desempenho dos Conjuntos de Classificadores das Séries 1 e 2.**

## Referências


Bhat, S. Y., Abulaish, M., and Mirza, A. A. (2014). Spammer classification using ensemble methods over structural social network features. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02*, pages 454–458. IEEE Computer Society.

Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C. I., and Kuncheva, L. I. (2015). Diversity techniques improve the performance of the best imbalance learning ensembles. volume 325, pages 98–117. Elsevier.

Easwaramoorthy, S., Thamburasa, S., Aravind, K., Bhushan, S. B., and Rajadurai, H. (2016). Heterogeneous classifier model for e-mail spam classification using fso feature selection method. In *Inventive Computation Technologies (ICICT), International Conference on*, volume 1, pages 1–6. IEEE.

Ibrahim, A. J., Siraj, M. M., and Din, M. M. (2017). Ensemble classifiers for spam review detection. In *Application, Information and Network Security (AINS), 2017 IEEE Conference on*, pages 130–134. IEEE.

Symantec (2018). Monthly threat report, Acessado em 20 de Junho de 2018. www.symantec.com/security_response/publications/monthlythreatreport.jsp.

Yin, X.-C., Huang, K., Hao, H.-W., Iqbal, K., and Wang, Z.-B. (2014). A novel classifier ensemble method with sparsity and diversity. In *Neurocomputing*, volume 134, pages 214–221. Elsevier.