Adriano Alonso Veloso

e-mail: adrianov@dcc.ufmg.br

Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

Classificação Associativa via Projeção de Treino

Construção de Classificadores a partir de Dados Discretos, Incertos, Dinâmicos e de Alta Dimensionalidade

Projeto apresentado como requisito à obtenção de Bolsa de Produtividade em Pesquisa do CNPq.

Sumário Executivo

Objetivos: O principal objetivo deste projeto é desenvolver métodos e algoritmos para a tarefa de classificação automática perante dados discretos, incertos, dinâmicos e de alta dimensionalidade.

Motivação: Dados discretos, incertos, dinâmicos e de alta dimensionalidade são pervasivos e produzidos pelas mais diversas aplicações. Texto livre ou documentos estruturados são exemplos de dados com tais características. Perante esse tipo de dado, a produção de modelos preditivos (i.e., classificadores) é uma tarefa extremamente desafiadora, seja pela alta complexidade computacional envolvida no processo, seja pela dificuldade em se atingir índices aceitáveis de efetividade e acurácia. A busca pela solução de tais desafios, bem como o impacto decorrente dos algoritmos a serem desenvolvidos, motivam este projeto.

Descrição: O projeto é organizado em quatro camadas. Na primeira camada iremos definir as fontes, coletar e preparar os dados com as características de nosso interesse. A segunda camada envolve a elaboração de novos métodos para produção de classificadores através do uso de projeções de treino. Na terceira camada serão desenvolvidos algoritmos de classificação baseados no método de projeção de treino. Finalmente, na quarta camada iremos avaliar os algoritmos desenvolvidos em diversas aplicações de reconhecida importância.

Resultados Esperados: Ao fim deste projeto (36 meses) espera-se obter os seguintes resultados: (i) projetar, implementar e validar novos métodos e algoritmos para classificação via projeção de treino; (ii) avaliar a efetividade prática dos algoritmos desenvolvidos em aplicações relevantes e desafiadoras; (iii) formar 2 doutores, 8 mestres e 10 alunos de iniciação científica; (iv) publicar 6 artigos em periódicos, 8 artigos em conferências internacionais e 8 artigos em conferências nacionais; e (v) publicar e transferir toda a tecnologia produzida durante o projeto para fins de pesquisa e desenvolvimento tecnológico.

Indicadores de Pesquisa: A seguir mencionamos os resultados de pesquisa obtidos nos últimos três anos (abordando temas relacionados ao projeto proposto). Publicamos um livro em 2011 pela Springer. Publicamos 8 artigos em periódicos (sendo que 4 deles foram recentemente aceitos para publicação), e 22 artigos em eventos (12 dos quais internacionais). Temos ainda 4 submissões para periódicos em avaliação. Formamos 2 alunos de iniciação científica (um deles foi premiado com o melhor trabalho de iniciação científica de 2011 pela Sociedade Brasileira de Computação). Formamos 2 mestres e mais 2 mestres devem ser formados até o final de 2011. Recebemos 6 premiações/distinções (5 nacionais e 1 internacional) e participamos de 9 projetos de pesquisa (4 deles como coordenador). Atuamos em comitês de programa (eventos nacionais e internacionais) e revisamos artigos para pelo menos uma dezena de periódicos. Uma relação detalhada de nossos resultados é apresentada no Anexo A.

Recursos Solicitados: Solicitamos uma bolsa de produtividade em pesquisa nível 2.

1 Introdução

A capacidade de modelar comportamentos e fenômenos com a finalidade de realizar tarefas preditivas vem tornando-se cada vez mais desejável nos mais diversos cenários. Um modelo preditivo consiste em uma combinação (semi-)ótima de vários atributos que mostram algum tipo de influência (ou correlação) no comportamento ou fenômeno sendo modelado. Na área de *marketing*, por exemplo, o sexo e a idade do cliente, bem com seu histórico de compras, podem ser atributos usados para modelar seu comportamento de forma a prever compras futuras.

Existem diversos tipos de tarefas baseadas em modelagem preditiva [56]. Especificamente, a tarefa que vamos abordar neste projeto é denominada *classificação* [30]. Nesse caso, um algoritmo de classificação analisa dados históricos de forma a realizar previsões sobre eventos futuros. Tais eventos são denominados *classes*, e devem ser especificados de antemão. Tipicamente os dados históricos fornecidos a um algoritmo de classificação são denominados *exemplos*, e cada exemplo é composto por um conjunto de atributos seguido da classe a qual o exemplo está associado. Os exemplos fornecidos ao algoritmo de classificação formam o *conjunto de treino*, e o algoritmo deve analisar os exemplos no conjunto de treino de forma a fornecer como saída um modelo (ou classificador), que tem a capacidade de predizer a classe que deve ser associada a dados futuros. Obviamente, o modelo produzido durante o processo de classificação pode cometer erros de predição. Sendo assim, o objetivo de um algoritmo de classificação é o de produzir um classificador que cometa a menor quantidade de erros possível quando submetido a dados futuros, obedecendo ainda a restrição de ser computacionalmente eficiente [59, 74, 75].

Neste projeto vamos assumir cenários de aplicação para algoritmos de classificação nos quais os dados a serem analisados possuem quatro características chaves:

1. São discretos: os atributos podem assumir um número finito de possíveis valores, os quais não podem ser sub-divididos.

2. São incertos: as classes e os atributos possuem um certo grau de incerteza, tendo em vista a possibilidade de erros na preparação e/ou coleta dos dados.

3. São dinâmicos: o conjunto de treino pode ser alterado, com a inclusão, remoção e alteração de exemplos.

4. Possuem alta dimensionalidade: a quantidade de atributos que compõem um exemplo é potencialmente enorme.

Algoritmos de classificação enfrentam sérios desafios ao lidar com dados discretos, incertos, dinâmicos e de alta dimensionalidade [48, 49]. Aqueles algoritmos que são especificamente desenvolvidos para analisar dados discretos estão seriamente sujeitos à explosão combinatória, o que os compromete do ponto de vista da eficiência computacional [40]. Dados incertos, por sua vez, podem comprometer a efetividade dos algoritmos de classificação devido à incerteza inerente aos valores assinalados às classes e aos atributos, o que pode afetar a qualidade do conjunto de treino [50]. Já a dinamicidade dos dados de treino obriga o algoritmo de classificação a ser altamente incremental [65]. Por fim, a alta dimensionalidade dos dados faz com que os algoritmos de classificação busquem

um meio-termo entre efetividade e eficiência, sendo que modelos (ou classificadores) mais efetivos não possam ser produzidos de forma eficiente [53, 47].

Mesmo com todas as dificuldades mencionadas acima, é extremamente importante produzir classificadores eficientes e eficazes a partir de dados discretos, incertos, dinâmicos e de alta dimensionalidade. Esse tipo de dado é cada vez mais comum devido à facilidade em produzí-lo e coletá-lo. Exemplos desse tipo de dado incluem texto livre, documentos estruturados, e até mesmo dados de microarranjo[1]. Entre os diversos cenários de aplicação que utilizam esse tipo de dado, podemos citar:

- Predição de desastres naturais, como enchentes e alagamentos.
- Predição de epidemias, como a da dengue [45].
- Descoberta de interações medicamentosas.
- Detecção de fraudes em transações eletrônicas.
- Detecção de conteúdo poluído na Web [10, 28].
- Deteção de objetos replicados.
- Extração de entidades [33, 34].
- Análise e monitoramento de sentimentos e opiniões [31, 65].
- Recomendação e ordenação de conteúdo e de produtos [2].
- Categorização e busca de conteúdo multimídia [38, 39].
- Manutenção de bibliotecas digitais [14, 41, 43].

No decorrer do projeto, iremos introduzir um método de classificação genérico, que habilita a produção de classificadores efetivos e eficazes perante dados discretos, incertos, dinâmicos e de alta dimensionalidade. Tal método é denominado *Classificação Associativa via Projeção de Treino*, e tem como intuição chave decompor um problema de classificação em sub-problemas menores, onde cada sub-problema pode ser solucionado efetivamente de maneira independente, e através da utilização de classificadores mais simplificados. A união da simplicidade e efetividade torna o método altamente prático, como discutido em [13].

Posteriormente, serão propostos diversos algoritmos de classificação baseados no método de classificação associativa via projeção de treino. Mais especificamente, apresentaremos estratégias desenvolvidas para lidar com dados discretos, dinâmicos, incertos e de alta dimensionalidade. Tais estratégias funcionam como "peças de montagem" de algoritmos de classificação. Dessa forma, apresentaremos um *pipeline* de montagem, que é dividido em 5 etapas:

[1] O termo microarranjo é uma tradução natural e aceita para o termo em Inglês *microarray* pelo qual uma técnica experimental da Biologia Molecular é mais conhecida.

1. Escolha da modalidade de classificação: o algoritmo de classificação pode adotar a modalidade supervisionada, semi-supervisionada ou ativa. Na modalidade supervisionada, o algoritmo de classificação terá acesso a um conjunto de treino composto por apenas exemplos para os quais as classes são conhecidas (i.e., exemplos rotulados). Na modalidade semi-supervisionada, o algoritmo de classificação terá acesso a um conjunto de treino composto tanto por exemplos rotulados quanto não-rotulados. Finalmente, na modalidade ativa, o algoritmo de classificação deverá ser capaz de selecionar os exemplos que irão compor o conjunto de treino.

2. Escolha da estratégia para lidar com dados discretos: este tipo de dado pode ser sumarizado por meio de padrões de co-ocorrência, e as estratégias disponíveis nesta etapa do pipeline diferem-se dependendo do padrão de co-ocorrência que irá compor o classificador.

3. Escolha da estratégia para lidar com dados dinâmicos: este tipo de dado exige que o algoritmo de classificação seja capaz de atualizar o classificador de forma rápida e constante, de forma a refletir alterações feitas no conjunto de treino. As estratégias disponíveis nesta etapa do pipeline diferem-se dependendo da frequência com que o conjunto de treino é modificado.

4. Escolha da estratégia para lidar com dados incertos: este tipo de dado contém imprecisões que podem comprometer a qualidade do conjunto de treino, exigindo que o algoritmo de classificação seja capaz de reduzir o grau de incerteza dos dados enquanto produz o classificador.

5. Escolha da estratégia para lidar com dados de alta dimensionalidade: este tipo de dado pode comprometer o desempenho computational do algoritmo de classificação, exigindo que o algoritmo de classificação adote cortes no espaço de busca por padrões ou que seja altamente paralelizável.

A grande diversidade de estratégias que podem ser escolhidas em cada etapa do pipeline resulta em mais de 100 possibilidades de algoritmos de classificação. Toda a interação entre dados, métodos, algoritmos e aplicações está ilustrada na Figura 6, que descreve a estrutura em camadas do projeto. Nas seções a seguir descreveremos melhor todos os conceitos mencionados bem como as frentes de pesquisa e atividades que compõem este projeto, mas antes discutimos a conformidade do projeto com os desafios de pesquisa definidos pela Sociedade Brasileira de Computação.

1.1 Desafios de Pesquisa em Computação

Em 2006, a Sociedade Brasileira de Computação promoveu o evento "Desafios de Pesquisa em Computação 2006-2016". Foram elencados cinco desafios de pesquisa, os quais dicutimos sob a perspectiva deste projeto.

O primeiro desafio é a gestão de informação em grandes volumes de dados multimídia distribuídos. Nesse sentido, podemos citar potenciais aplicações para nossos algoritmos, como a busca e a ordenação de imagens [38, 39], e a recomendação de conteúdo multimídia [54], como músicas e vídeos.

O segundo desafio é a modelagem computacional de sistemas complexos de várias naturezas. Nossos algoritmos podem ser usados para modelar sistemas altamente complexos (ou partes deles), tais como máquinas de busca [2], sistemas de vigilância de epidemias [45] e desastres naturais, monitores de sentimentos e opiniões [67, 63, 65, 11, 71], ou bibliotecas digitais [43, 41, 19]. Todos esses sistemas, e outros mais, podem ser beneficiados por algoritmos de classificação que seguem o método proposto no projeto: classificação associativa via projeção de treino.

O terceiro desafio está relacionado ao impacto das novas tecnologias de processamento como computação biológica. De fato, nossos algoritmos podem auxiliar diversas aplicações ligadas à bioinformática, tais como a detecção de proteínas homólogas [9], ou a descoberta de interações medicamentosas e enzimáticas.

O quarto desafio tem por objetivo promover o acesso participativo e universal do cidadão brasileiro ao conhecimento. Nesse sentido, nossos algoritmos vêm sendo utilizados no contexto do projeto "Observatório da Web[2]" – onde auxiliam ferramentas gratuitas dedicadas ao monitoramento de importantes fatos, eventos e entidades na rede mundial de computadores em tempo real.

Finalmente, o quinto desafio tem por objetivo desenvolver sistemas onivalentes. Tais sistemas estarão presentes nos mais diversos ambientes e atividades humanas, prestando serviços essenciais para a saúde, educação, informação, comunicação etc. Tarefas de aprendizado de máquina e modelagem preditiva, como a tarefa de classificação, estão cada vez mais acopladas a aparelhos domésticos [55], automóveis [51, 72], aviões [37], salas de cirurgia [29, 68], simuladores de vôo e de batalhas aéreas [70] etc.

Desta forma, acreditamos que o nosso projeto está em conformidade e irá contribuir efetivamente na busca pela solução dos desafios de pesquisa elencados pela academia brasileira na área de Ciência da Computação.

1.2 Objetivos Específicos

Os principais objetivos deste projeto são:

- Aprimorar e estender o método de classificação associativa via projeção de treino.
- Desenvolver estratégias para lidar com os desafios impostos por dados discretos, dinâmicos, incertos e de alta dimensionalidade.
- Desenvolver uma metodologia para combinar as diversas estratégias desenvolvidas, o que resultará em diversos algoritmos de classificação.

2 Descrição do Projeto

Este projeto de pesquisa versa sobre a elaboração, desenvolvimento e avaliação de novos métodos e algoritmos de classificação que produzam classificadores eficazes de forma eficiente, quando submetidos a dados discretos, incertos, dinâmicos, e de alta dimensionalidade. A escolha por atuar em dados com tais características aconteceu em virtude

[2] `http://www.observatorio.inweb.org.br/`

de alguns fatores. O primeiro deles é a onipresença de dados com alguma dessas características nas mais diversas aplicações, tendo em vista a atual facilidade em produzí-los, coletá-los e armazená-los (i.e., texto livre ou estruturado). O segundo fator é o desafio de desenvolver algoritmos de classificação eficazes e eficientes perante dados com tais características. Por último, destacamos a grande demanda por algoritmos de classificação que atuem sob tal tipo de dado, como evidenciado pelo grande número de aplicações nas quais esses algoritmos podem contribuir.

2.1 Classificação

Para entendermos melhor os conceitos ligados à tarefa de classificação, apresentamos alguns exemplos:

Exemplo 1 Definimos crédito como sendo uma quantidade de dinheiro emprestada por uma instituição financeira, geralmente um banco, e que deve ser posteriormente paga com o acréscimo de juros. É importante para o banco identificar de antemão o risco associado ao empréstimo, que pode ser visto como sendo a probabilidade do cliente não pagá-lo. Isso é importante tanto para protejer o lucro do banco, quanto para evitar com que o cliente faça empréstimos que estão além de sua capacidade de pagamento.

Geralmente o banco calcula o risco associado a um cliente com base na quantidade desejada de crédito e em atributos do cliente, tais como salário, poupança, profissão, idade etc. O banco armazena registros históricos de empréstimos passados realizados por outros clientes, juntamente com a informação sobre a quitação (ou não) do empréstimo. A partir desses registros, o banco busca inferir um conjunto de regras que codificam a associação entre os atributos de um cliente e o risco associado a ele. Tais regras formam um modelo preditivo que chamamos de classificador, e o banco pode usá-lo para decidir aceitar ou recusar um pedido de empréstimo.

Após analisar o conjunto de registros representando empréstimos pagos e não-pagos, o algoritmo de classificação deve produzir regras da forma:

$$\text{SE salário} \geq \theta_1 \text{ E empréstimo} \leq \theta_2 \text{ ENTÃO o risco é BAIXO}$$

Os valores para θ_1 e θ_2 devem ser escolhidos adequadamente pelo algoritmo de classificação de forma a produzir o classificador (veja Figura 1), e tal escolha é encarada como uma combinação (semi-)ótima dos atributos. Regras como essa são usadas pricipalmente com a finalidade de predição, assumindo que o fenômeno que produziu os dados passados é o mesmo fenômeno que produzirá os dados futuros.

Exemplo 2 A interação entre diferentes fármacos (ou medicamentos) é um evento clínico em que os efeitos de um fármaco são alterados pela presença de outro. Essa interação constitui causa comum de efeitos adversos. Quando dois medicamentos são administrados, concomitantemente, a um paciente, eles podem agir de forma independente ou interagirem entre si, com aumento ou diminuição de efeito terapêutico ou tóxico

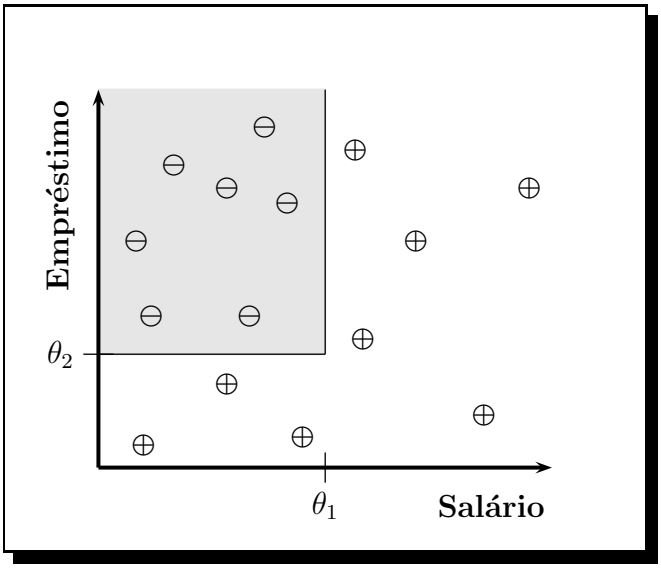

Figura 1: O símbolo $\ominus$ representa um registro no qual o empréstimo não foi pago. O símbolo $\oplus$ representa um registro no qual o empréstimo foi pago.

de um ou de outro. O desfecho de uma interação medicamentosa pode ser perigoso quando promove aumento da toxicidade de um fármaco. Além disso, algumas vezes a interação medicamentosa reduz a eficácia de um fármaco, o que também pode ser altamente nocivo. Por fim, há interações que podem ser benéficas e muito úteis, como na co-prescrição deliberada de anti-hipertensivos e diuréticos.

Uma maneira de modelar as interações medicamentosas é através da utilização de grafos bipartidos. Nesse caso, o nome do medicamento é ligado a características dele, por exemplo as palavras que aparecem em sua bula (veja Figura 2). Outras características, como a expansão enzimática ou o princípio ativo do medicamento, podem ser exploradas também. Seja qual for o caso, é esperado que diferentes medicamentos compartilhem características em comum. Dessa forma, dado um conjunto de treino, onde cada exemplo representa as características de um par de medicamentos, o algoritmo de classificação deve encontrar padrões de interações graves, moderadas ou leves, e utilizar esses padrões de forma a produzir um classificador, que será capaz de predizer a gravidade de interações medicamentosas ainda desconhecidas.

Exemplo 3 Pagamentos feitos pela internet por meio de cartões de crédito são operações cada vez mais comuns em todo mundo. De forma a aumentar a segurança nesse tipo de transação, algumas empresas oferecem o serviço de intermediação entre compradores e vendedores. Nesse caso, o comprador tem a garantia de produto ou serviço entregue (ou dinheiro de volta), enquanto o vendedor fica livre de perdas em suas vendas. O intermediador cobra uma porcentagem sobre as transações realizadas com sucesso, mas arca com os custos decorrentes de fraudes. Sendo assim, é do interesse do intermediador detectar transações potencialmente fraudulentas, de forma a invalidá-las ou cancelá-las com o intuito de maximizar os lucros.

Milhares de transações são armazenadas todos os dias, à medida em que elas vão

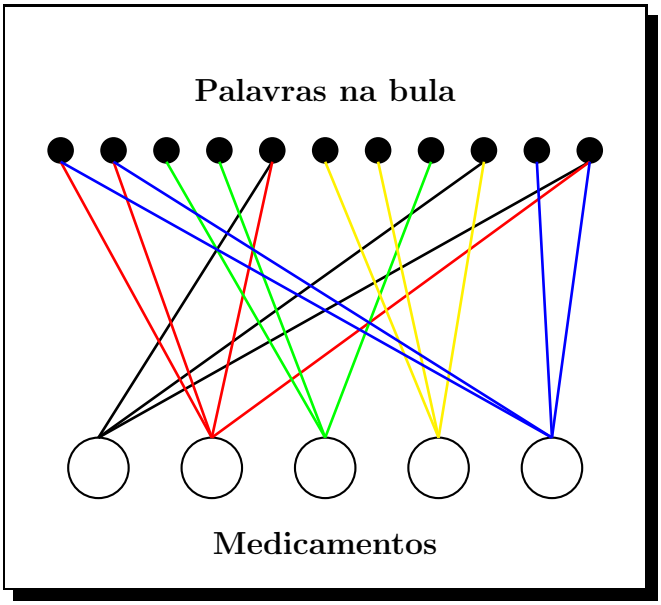

Figura 2: Grafo bipartido representando interações medicamentosas.

sendo concluídas. Cada transação consiste em uma série de atributos com informações sobre o comprador, vendedor, produto etc. Dessa forma, temos disponível um conjunto de exemplos de transações legítimas ou fraudulentas, e podemos assim utilizar algoritmos de classificação para, por exemplo, ordenar as transações de acordo com o potencial de fraude.

Contudo, novos padrões de fraude são constantemente criados, em uma tentativa de burlar sistemas de detecção. Sendo assim, de forma a detectar rapidamente novos perfis de fraude ainda desconhecidos, torna-se necessário manter o classificador coerente com transações recentes. No entanto, a enorme quantidade de transações disponíveis para treino inviabiliza a constante construção de classificadores. Uma alternativa mais eficiente é manter o classificador atualizado de maneira incremental, ou seja, ao invés de produzir outro classificador, o algoritmo de classificação atualiza o classificador de forma a refletir as novas transações. O resultado final é o mesmo classificador que seria produzido levando-se em conta todas as transações disponíveis para treino, porém o tempo necessário para atualizar o classificador corresponde à uma pequena fração do tempo necessário para produzí-lo novamente.

Com o classificador constantemente atualizado, o intermediador pode usá-lo para ordenar as transações, de forma a concentrar esforços de análise manual naquelas com maior indício de fraude (veja Figura 3).

Embora tenham sido discutidos superficialmente, os exemplos acima nos permitem observar alguns fatores importantes. O primeiro é que as soluções para diversos problemas podem ser modeladas através da tarefa de classificação. Para tanto, classificadores podem ser utilizados de maneiras diferentes (detecção, ordenação, identificação etc.). Além disso, diferentes problemas podem demandar diferentes características de um algoritmo de classificação, dependendo do tipo de dados envolvidos na análise e nas restrições impostas pelo problema a ser solucionado. Nesse sentido, ao processar dados dinâmicos, o algoritmo

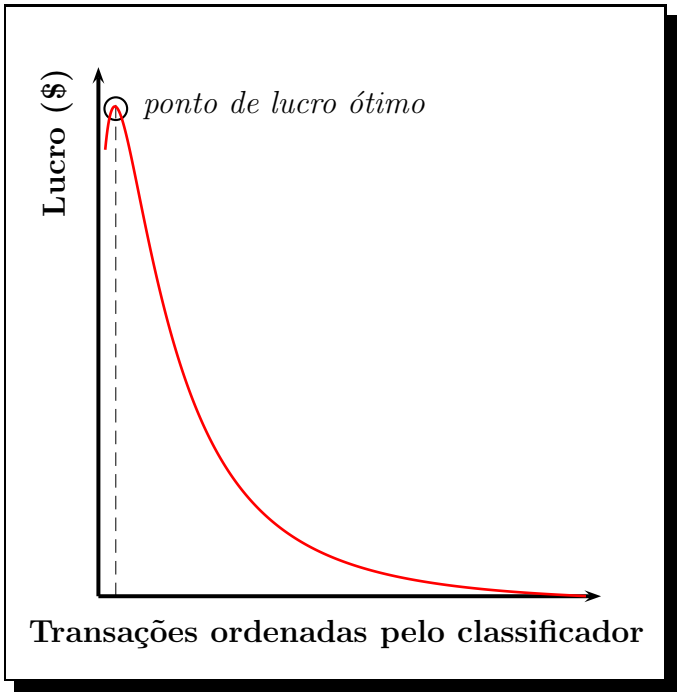

Figura 3: Transações são ordenadas, de forma a posicionar aquelas com maior propensão de fraude no início do eixo-x.

de classificação deve ter a capacidade de atualizar o classificador de maneira incremental. Ao processar dados de alta dimensionalidade, o algoritmo de classificação deve ter a capacidade de selecionar apenas os atributos relevantes, sem que haja perda de informação e ao mesmo tempo evitando os efeitos indesejados da explosão combinatória. Claramente, existem outras características e restrições, e portanto torna-se necessário um método geral o suficiente que consiga prover os subsídios necessários que habilitem o desenvolvimento de algoritmos de classificação que sejam eficazes e eficientes perante diferentes restrições.

2.2 Atividades de Pesquisa

As atividades de pesquisa (e algumas de implementação) associadas a este projeto podem acontecer em quatro camadas, discutidas a seguir:

Camada 1: Dados

Duas operações precisam ser realizadas antes de podermos utilizar nossos métodos e algoritmos de classificação. Discutimos estas operações a seguir.

- Coleta: neste projeto trabalharemos com dados provenientes de diversas fontes, incluindo texto-livre extraído de mídias sociais como Twitter[3] e Facebook[4], metadados extraídos de bibliotecas digitais como a DBLP[5], dados de referência dispo-

[3] `www.twitter.com`
[4] `www.facebook.com`
[5] `www.informatik.uni-trier.de/~ley/db`

nibilizados publicamente[6,7,8], dados obtidos através de parcerias e convênios com empresas públicas, privadas e agências governamentais. As atividades referentes à coleta vão desde o desenvolvimento de técnicas de extração dos dados e amostragem [46], até a anonimização de dados sensíveis.

- Rotulação: como mencionado anteriormente, o algoritmo de classificação necessita de um conjunto de treino, composto por exemplos. Idealmente, cada exemplo consiste de uma série de atributos, que são acompanhados de uma (ou mais) classe(s). Chamamos de rotulação o ato de associar uma (ou mais) classe(s) aos exemplos. Algumas fontes já fornecem dados rotulados. Outras fontes, porém, fornecem apenas os atributos associados aos dados, e a informação de classe precisa ser produzida por meio de rotulação. A rotulação pode ser realizada de forma manual ou semi-automática. No caso da rotulação manual, voluntários inspecionam um conjunto reduzido de exemplos, e associam a(s) classe(s) que acharem ser mais plausíveis a esses exemplos. Geralmente observa-se uma discordância entre os voluntários, e por isso os exemplos produzidos possuem um certo grau de incerteza. Em algumas circunstâncias é possível realizar a rotulação semi-automática. Nesse caso, explora-se algum tipo de informação mais robusta associada ao exemplos [31], como *hashtags*, viés de usuários etc. A rotulação semi-automática permite a criação de conjuntos de treino maiores do que os produzidos pela rotulação manual, mas a incerteza associada aos exemplos também é maior.

Camada 2: Métodos

A seguir detalhamos os conceitos sobre os métodos baseados em classificação associativa via projeção de treino, que serão adotados pelos algoritmos de classificação a serem propostos neste projeto.

Treino e teste – em um problema de classificação, a entrada é composta por um conjunto de tuplas (ou exemplos) com a forma $r_i = (x_i, y_i)$. Cada x_i é representado como um registro de tamanho fixo (n atributos) com a forma $< a_1, a_2, \ldots, a_n >$, onde a_k é o valor associado ao atributo k. Cada y_i pode conter valores de um conjunto finito de possibilidades $\{c_1, c_2, \ldots, c_m\}$, e indica a classe para a qual r_i está associado. Casos para os quais $y_i =?$ indicam que a classe de r_i é desconhecida. Geralmente assume-se que exista uma distribuição de probabilidade desconhecida $\mathrm{P}(y|x)$, ou seja, os valores dos atributos e as classes são relacionados de alguma forma. O conjunto de tuplas é dividido em duas partições, o conjunto de treino (denotado por $\mathcal{S}$) e o conjunto de teste (denotado por $\mathcal{T}$), da seguinte forma:

$$\begin{aligned} \mathcal{S} &= \{(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)\} \\ \mathcal{T} &= \{(x_{p+1}, ?), (x_{p+2}, ?), \ldots, (x_{p+q}, ?)\} \end{aligned}$$

[6] `www.netflix.com`
[7] `research.microsoft.com/en-us/um/beijing/projects/letor/`
[8] `www.drugbank.ca`

A Tabela 1 mostra a entrada para o problema de predição do risco de crédito, discutido anteriormente na Seção 2.1. Na tabela temos um conjunto de treino composto por 15 exemplos, enquanto o conjunto (por simplicidade), é composto por apenas um exemplo.

Tabela 1: Entrada para um problema de classificação.

	x_i		
	Salário	Empréstimo	y_i
x_1	menor que R$1.000	maior que R$500	$\ominus$
x_2	entre R$1.000 e R$2.000	maior que R$500	$\ominus$
x_3	entre R$1.000 e R$2.000	maior que R$500	$\ominus$
x_4	menor que R$1.000	maior que R$500	$\ominus$
x_5	entre R$1.000 e R$2.000	maior que R$500	$\ominus$
x_6	menor que R$1.000	maior que R$500	$\ominus$
x_7	entre R$1.000 e R$2.000	maior que R$500	$\ominus$
x_8	entre R$1.000 e R$2.000	maior que R$500	$\oplus$
x_9	maior que R$2.000	maior que R$500	$\oplus$
x_{10}	menor que R$1.000	menor que R$500	$\oplus$
x_{11}	maior que R$2.000	menor que R$500	$\oplus$
x_{12}	maior que R$2.000	menor que R$500	$\oplus$
x_{13}	entre R$1.000 e R$2.000	menor que R$500	$\oplus$
x_{14}	maior que R$2.000	maior que R$500	$\oplus$
x_{15}	maior que R$2.000	maior que R$500	$\oplus$
x_{16}	menor que R$1.000	maior que R$500	?

Regras de associação – Os métodos de classificação a serem propostos produzem classificadores compostos por estruturas chamadas *regras de associação* [1]. Especificamente em nosso contexto, essas regras tem a forma $\{\mathcal{X} \rightarrow c_j\}$, onde $\mathcal{X}$ é uma combinação qualquer de atributos[9], e c_j é uma classe. Logo, essas regras podem ser consideradas como sendo um mapeamento de atributos para classes, de forma a aproximar $\mathrm{P}(y|x)$. Regras de associação são extraídas dos exemplos em $\mathcal{S}$, e cada regra possui três propriedades importantes:

- Suporte: o suporte de uma regra $\{\mathcal{X} \rightarrow c_j\}$, que é denotado por $\sigma(\mathcal{X} \rightarrow c_j)$, é um indicativo importante sobre a confiabilidade da associação entre $\mathcal{X}$ e c_j, e é calculado como a fração de exemplos em $\mathcal{S}$ contendo $\mathcal{X}$ como subconjunto, para os quais a classe associada é c_j. Formalmente:

$$\sigma(\mathcal{X} \rightarrow c_j) = \frac{|(x_i, y_i)| \in \mathcal{S} \text{ tal que } \mathcal{X} \subseteq x_i \text{ e } c_j = y_i}{|\mathcal{S}|}$$

- Confiança: a confiança de uma regra $\{\mathcal{X} \rightarrow c_j\}$, que é denotada por $\theta(\mathcal{X} \rightarrow c_j)$, é um indicativo importante sobre o nível da associação entre $\mathcal{X}$ e c_j, e é calculada

[9]Sempre nos referimos a atributo como sendo o valor associado ao atributo.

como a fração de exemplos em $\mathcal{S}$ contendo $\mathcal{X}$ como subconjunto, para os quais a classe correspondente é c_j. Formalmente:

$$\theta(\mathcal{X} \rightarrow c_j) = \frac{|(x_i, y_i)| \in \mathcal{S} \text{ tal que } \mathcal{X} \subseteq x_i \text{ e } c_j = y_i}{|(x_i, y_i)| \in \mathcal{S} \text{ tal que } \mathcal{X} \subseteq x_i}$$

- Utilidade: dizemos que uma regra $\{\mathcal{X} \rightarrow c_j\}$ é útil para classificar um exemplo (x_i, y_i) se e somente se $\mathcal{X} \subseteq x_i$. Uma regra é útil se ela for útil para pelo menos um exemplo em $\mathcal{T}$.

Projeção que preserva correlação – A característica mais marcante de nossos métodos de classificação é a utilização de projeções do treino. Mais especificamente, toda vez que um exemplo $(x_i, ?) \in \mathcal{T}$ é submetido ao classificador, os atributos em x_i são utilizados como um filtro que remove de $\mathcal{S}$ todos os atributos que não estão em x_i. Este procedimento resulta em uma projeção de $\mathcal{S}$, denotada por $\mathcal{S}^{x_i}$. A Tabela 2 mostra a projeção para o exemplo x_{16}. Após construir a projeção, as regras são extraídas a partir de $\mathcal{S}^{x_i}$. Como mostrado em [22] a utilização de projeções não deforma (nem altera) as correlações entre atributos e classes, e ao mesmo tempo, assegura que todas as regras extraídas sejam úteis.

Tabela 2: Projeção para x_{16}.

	x_i		
	Salário	Empréstimo	y_i
x_1	menor que R$1.000	maior que R$500	$\ominus$
x_2		maior que R$500	$\ominus$
x_3		maior que R$500	$\ominus$
x_4	menor que R$1.000	maior que R$500	$\ominus$
x_5		maior que R$500	$\ominus$
x_6	menor que R$1.000	maior que R$500	$\ominus$
x_7		maior que R$500	$\ominus$
x_8		maior que R$500	$\oplus$
x_9		maior que R$500	$\oplus$
x_{10}	menor que R$1.000		$\oplus$
x_{14}		maior que R$500	$\oplus$
x_{15}		maior que R$500	$\oplus$

Prova de eficiência e efetividade – Qualquer algoritmo de classificação que seja baseado nos métodos de classificação associativa via projeção de treino é asseguradamente eficiente e efetivo, no sentido de que:

- O número necessário de exemplos para que o algoritmo consiga produzir um classificador com erro empírico[10] próximo de 0, cresce polinomialmente com o número de atributos.

[10]Quantidade de erros cometidos pelo classificador quando este prediz as classes dos exemplos no próprio conjunto de treino.

- O algoritmo consegue produzir um classificador em tempo polinomial.

► **Desenvolvemos provas de eficiência e efetividade em** [13, 54, 65]. Tais provas demonstram que o método de classificação associativa via projeção de treino é na verdade um habilitador para a produção de classificadores perante dados discretos, incertos, dinâmicos e de alta dimensionalidade.

Aproximação de probabilidades – Denotamos o conjunto de regras extraídas da projeção $\mathcal{S}^{x_i}$ por $\mathcal{R}^{x_i}$. Além disso, denotamos o subconjunto de regras em $\mathcal{S}^{x_i}$ que predizem a classe c_j por $\mathcal{R}^{x_i}_{c_j}$. Cada regra em $\mathcal{R}^{x_i}_{c_j}$ é interpretada como um voto para a classe c_j. Esses votos (i.e., regras) podem ser ponderados de diferentes formas.

- Confiança média: dado em exemplo $(x_i, ?) \in \mathcal{T}$, calcula-se uma pontuação para cada classe em $\{c_1, c_2, \ldots, c_m\}$, conforme mostrado na Equação 1. Em seguida as pontuações são normalizadas, fornecendo a probabilidade de pertinência associada à cada classe, conforme mostra a Equação 2.

$$s(x_i, c_j) = \frac{\sum_{r \in \mathcal{R}^{x_i}_{c_j}} \theta(r)}{\mid \mathcal{R}^{x_i}_{c_j} \mid} \tag{1}$$

$$\hat{p}(c_j|x_i) = \frac{s(x_i, c_j)}{\sum s(x_i, c_k)} \tag{2}$$

- Ponderação otimizada de regras: dado em exemplo $(x_i, ?) \in \mathcal{T}$, calcula-se a probabilidade de pertinência associada à cada classe através de métricas baseadas no erro empírico [26] das regras em $\mathcal{R}^{x_i}$.

Camada 3: Pipeline para a Montagem de Algoritmos

A seguir apresentamos diversas estratégias desenvolvidas para lidar com dados discretos, dinâmicos, incertos e de alta dimensionalidade. Essas estratégias podem ser vistas como peças a serem usadas na montagem de um algoritmo de classificação. Sendo assim, temos uma gama de possíveis algoritmos de classificação que diferem-se dependendo das estratégias (peças) adotadas. Além disso, os algoritmos também podem adotar diferentes modalidades de classificação (ou tipos de aprendizado). Ao final do pipeline de montagem, temos um novo algoritmo de classificação, que é na verdade uma das várias possíveis instanciações do método de classificação associativa via projeção de treino.

Modalidades de classificação – A primeira etapa do pipeline de montagem do algoritmo de classificação consiste da escolha da modalidade de classificação. Assumiremos três possibilidades:

- Modalidade supervisionada (M_1): nesta modalidade o algoritmo de classificação produz um classificador a partir de um conjunto de treino composto apenas por exemplos rotulados [10, 14].

- Modalidade semi-supervisionada (M_2): nesta modalidade o algoritmo de classificação produz um classificador a partir de um conjunto de treino composto por poucos exemplos rotulados e muitos exemplos não-rotulados [33].
- Modalidade ativa (M_3): nesta modalidade o algoritmo de classificação deve selecionar os exemplos mais relevantes para formar o conjunto de treino, a partir do qual o classificador será produzido [35, 69].

A escolha da modalidade de classificação deve levar em conta o custo associado à rotulação do conjunto de treino.

Estratégias para lidar com dados discretos – A segunda etapa do pipeline de montagem do algoritmo de classificação consiste da escolha do tipo de padrão de co-ocorrência que irá compor o classificador. Assumiremos quatro possibilidades:

- Regras baseadas em padrões simples (a_1): o algoritmo de classificação produz um classificador composto por regras do tipo $\{\mathcal{X} \rightarrow c_j\}$, onde $\mathcal{X}$ é qualquer combinação de atributos [3, 52, 73].
- Regras baseadas em padrões fechados (a_2): o algoritmo de classificação produz um o classificador composto por regras do tipo $\{\mathcal{X} \rightarrow c_j\}$, sendo que que não existe outra regra $\{\mathcal{Y} \rightarrow c_j\}$ tal que $\mathcal{X} \subseteq \mathcal{Y}$ e $\sigma(\mathcal{X} \rightarrow c_y) = \sigma(\mathcal{Y} \rightarrow c_j)$. Os padrões fechados formam um subconjunto dos padrões simples.
- Regras baseadas em padrões maximais (a_3): o algoritmo de classificação produz um classificador composto por regras do tipo $\{\mathcal{X} \rightarrow c_j\}$, sendo que que não existe outra regra $\{\mathcal{Y} \rightarrow c_j\}$ tal que $\mathcal{X} \subseteq \mathcal{Y}$. Os padrões maximais formam um subconjunto dos padrões fechados.
- Regras baseadas em padrões sequenciais (a_4): o algoritmo de classificação produz um classificador composto por regras do tipo $\{\mathcal{X} \Rightarrow \mathcal{Y} \Rightarrow \ldots \Rightarrow \mathcal{Z} \rightarrow c_j\}$, onde $\{a \Rightarrow b\}$ indica a precedência temporal do padrão a sobre o padrão b.

A escolha da estratégia para lidar com dados discretos deve levar em conta a quantidade de padrões que podem ser enumerados, o nível de redundância de informação aceitável (i.e., redução de redundância pela remoção de subconjuntos), e a existência de uma ordenação temporal entre os exemplos [32].

Estratégias para lidar com dados dinâmicos – A terceira etapa do pipeline de montagem do algoritmo de classificação consiste da escolha da estratégia de atualização do classificador. Assumiremos duas possibilidades:

- Atualização incremental (b_1): o algoritmo de classificação produz um classificador que atualiza as regras com base nos dados mais recentes [4, 10, 15, 16, 25, 57, 58].
- Processamento em janela deslizante (b_2): o algoritmo de classificação produz um classificador com base nos dados mais recentes e desconsiderando os dados mais antigos [65, 66].

A escolha pela estratégia para lidar com dados dinâmicos deve levar em conta o impacto de dados recentes e antigos na efetividade do classificador.

Estratégias para lidar com dados incertos – A quarta etapa do pipeline de montagem do algoritmo de classificação consiste da escolha da estratégia de redução de incerteza dos dados. Assumiremos três possibilidades:

- Maximização de probabilidade (*expectation maximization* [36]) (c_1): o algoritmo de classificação produz um classificador de maneira iterativa, onde em cada iteração o conjunto de treino é modificado de forma a convergir para uma configuração com menos incerteza [34].
- Calibração (c_2): o algoritmo de classificação produz um classificador calibrado, ou seja, as probabilidades $\hat{p}(c_j|x_i)$ são corrigidas de forma a minimizar efeitos negativos decorrentes da distribuição acentuada das classes no conjunto de treino [19, 23, 26].
- Combinação e agregação de visões e listas (c_3): o algoritmo de classificação produz um classificador através da agregação de diferentes visões do conjunto de treino [7, 61]. Ou seja, há um particionamento vertical do conjunto de treino, de forma a minimizar os efeitos decorrentes da incerteza dos dados através da diversidade de fontes e visões.

A escolha pela estratégia para lidar com dados incertos deve levar em conta o nível de incerteza dos dados, bem como se a fonte de incerteza encontra-se nos atributos ou nas classes (ou em ambos).

Estratégias para lidar com dados de alta dimensionalidade – A quinta etapa do pipeline de montagem do algoritmo de classificação consiste da escolha da estratégia de enumeração de padrões ou de aumento de escalabilidade. Assumiremos duas possibilidades:

- Combinação por proximidade de atributos (d_1): o algoritmo de classificação evita a explosão combinatorial realizando combinações apenas entre atributos próximos. Dessa forma, a quantidade de combinações é drasticamente reduzida. O conceito de proximidade entre atributos (ou termos) é bastante aceito para o processamento de texto livre ou estruturado.
- Paralelização (d_2): o algoritmo de classificação deve ser altamente paralelizável [18, 20, 24].

Camada 4: Cenários de Aplicação

A seguir elencamos algumas aplicações que produzem dados discretos, dinâmicos, incertos e de alta dimensionalidade, e que portanto serão beneficiadas com nossos algoritmos de classificação.

- Predição de desastres naturais e epidemias: o classificador deverá processar/filtrar dados históricos usados que são correlacionados com incidências de dengue e número de alagamentos. Por exemplo, dado um conjunto de sentenças textuais coletadas de *microblogs* e mídias sociais, o classificador deverá filtrar apenas as sentenças que mencionam alguma experiência pessoal com dengue, ou apenas sentenças que possuem tempo verbal no presente.

- Descoberta de interações medicamentosas: o classificador deverá predizer o nível de gravidade na interação entre dois fármacos, que pode ser leve, moderada ou grave.
- Detecção de fraudes eletrônicas: o classificador deverá predizer a legitimidade de uma compra ou um pagamento feito pela internet via cartão de crédito.
- Detecção de conteúdo poluído na Web: o classificador deverá predizer se um determinado conteúdo Web (i.e., vídeo, comentário etc.) foi postado com finalidade maliciosa ou oportunista [28].
- Detecção de objetos replicados: o classificador deverá predizer se um determinado objeto (i.e., *websites*, páginas, imagens etc.) é replicado ou não.
- Desambiguação de entidades: o classificador deverá predizer a entidade referida em um determinado objeto (i.e., autores em citações [6, 42], empresas ou times de futebol em sentenças postas em *microblogs* [33] etc.).
- Análise e monitoramento de sentimentos: o classificador deverá predizer a opinião (ou o sentimento) associado a sentenças postadas em *microblogs* e blogs [64, 44].
- Recomendação de produtos e conteúdo: o classificador deverá predizer produtos associados a certos conteúdos (i.e., livros ou vídeos associados a *tags* [54]).
- Categorização e busca de imagens: o classificador deverá facilitar a busca e a categorização de conteúdo multimídia, como vídeos e imagens [39].
- Manutenção de bibliotecas digitais de citações: o classificador deverá reduzir a ambiguidade entre autores de citações que são armazenadas em bibliotecas digitais [41, 43], facilitando assim sua manutenção.
- Ordenação de documentos Web: o classificador deverá predizer a relevância de documentos retornados por máquinas de busca [8].

3 Metodologia do Projeto

Nesta seção discutimos a metodologia (ilustrada na Figura 6) que adotaremos durante o desenvolvimento do projeto proposto.

3.1 Coleta e Rotulação dos Dados

A coleta dos dados provenientes de mídias sociais e *blogs* está sendo realizada no laboratório SPEED[11] (Departamento de Ciência da Computação da UFMG). Diariamente coletamos aproximadamente 40MB de texto livre. Quando possível, uma parte dos dados é rotulada de forma semi-automática. A outra parte dos dados é amostrada, e essa amostra é rotulada de forma manual por voluntários [33], que podem ser estudantes do laboratório ou especialistas. Dados serão rotulados por pelo menos três voluntários, de

[11]Systems Performance Evaluation and Experimental Development

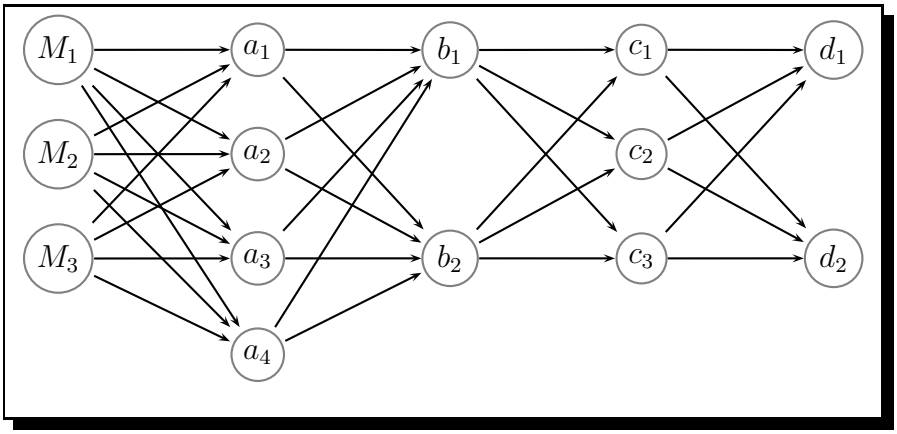

Figura 4: Interação entre modalidades de classificação (M_1, M_2, M_3) e estratégias para lidar com dados discretos (a_1, a_2, a_3, a_4), dinâmicos (b_1, b_2), incertos (c_1, c_2, c_3) e de alta dimensionalidade (d_1, d_2). Cada caminho partindo de M_1, M_2 ou M_3, até d_1 ou d_2, define um algoritmo de classificação baseado no método de classificação associativa via projeção de treino.

forma a reduzir o nível de subjetividade nas avaliações. Após rotulados, os dados serão usados não apenas para fins de treino, como também para fins de gabarito.

Dados de referência e dados oficiais são obtidos por meio de parcerias. Geralmente esses dados são estruturados e já são fornecidos de forma rotulada. Estes dados são utilizados para confrontar ou validar nossos resultados. Por exemplo, utilizamos dados oficiais do Ministério da Saúde para validar o monitoramento de epidemias através do Twitter[12], ou utilizamos dados de um grande portal Brasileiro para validar a aplicação de nossos classificadores para detecção de fraudes eletrônicas.

3.2 Elaboração de Métodos de Classificação via Projeção

Os métodos de classificação associativa via projeção de treino serão desenvolvidos e implementados por alunos sob a orientação do proponente. Após implementados, os métodos serão disponibilizados por meio de código aberto de alta qualidade, de forma que outros grupos de pesquisa possam também desenvolver algoritmos de classificação baseados no método de classificação associativa via projeção de treino.

3.3 Desenvolvimento e Implementação de Algoritmos

As estratégias que alimentarão o pipeline de montagem de algoritmos de classificação serão desenvolvidas e implementadas por alunos sob orientação do proponente. Os novos algoritmos que utilizarem as estratégias que compõem o pipeline serão o cerne de dissertações e teses.

O pipeline de montagem seguirá um protocolo de forma a facilitar a comunicação entre as diversas etapas. Sendo assim, estratégias associadas à uma mesma etapa do pipeline

[12] `http://www.newscientist.com/article/mg21128215.600-twitter-to-track-dengue-fever-outbreaks-in-brazil.html`

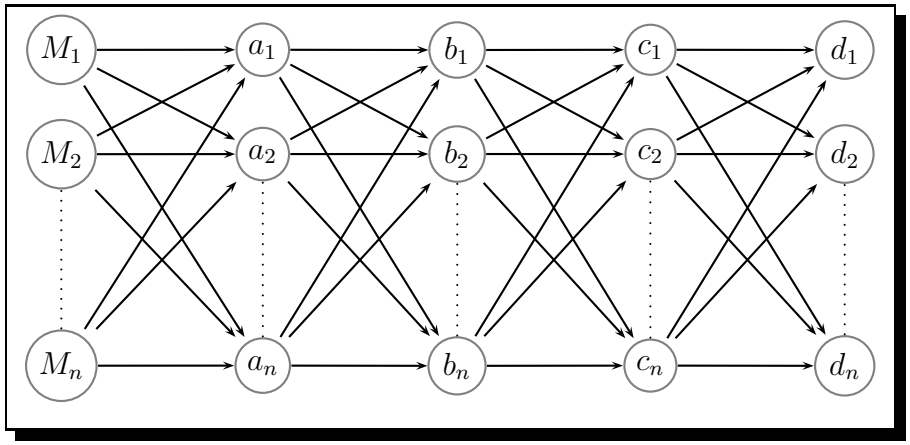

Figura 5: Interação entre modalidades de classificação (M_i's) e estratégias para lidar com dados discretos (a_i's), dinâmicos (b_i's), incertos (c_i's) e de alta dimensionalidade (d_i's).

serão implementadas de forma a obedecer o mesmo padrão de entrada e saída. Tal protocolo será muito importante já que o ambiente de desenvolvimento será compartilhado por diversos alunos, que irão interagir e compartilhar implementações da forma mais transparente possível. A Figura 4 ilustra o pipeline de montagem de algoritmos de classificação. Nesse caso, a implementação de uma estratégia deverá obedecer um protocolo de entrada de parâmetros e saída de valores. O protocolo irá facilitar também a participação de outros grupos de pesquisa interessados no desenvolvimento de algoritmos de classificação que atuem perante dados discretos, dinâmicos, incertos e de alta dimensionalidade. Nesse caso as implementações podem ser compartilhadas em uma escala maior, como ilustrado na Figura 5.

3.4 Avaliação e Cenários de Aplicação

Cada algoritmo de classificação corresponderá à uma instanciação do método de classificação associativa via projeção de treino, e será avaliado em aplicações relevantes. As avaliações experimentais serão objeto do estudo de alunos de iniciação científica sob a orientação do proponente.

Diversos cenários de aplicação serão empregados para fins de avaliação. Para tanto, formamos parcerias com empresas privadas e públicas, bem como com parcerias com orgãos governamentais. Para viabilizar a análise dos resultados, também formamos parceria com especialistas da Faculdade de Farmácia da UFMG, e com o Departamento de Bioquímica e Imunologia da UFMG. Finalmente, formamos cooperações com outros professores dos Departamentos de Ciência da Computação da UFMG, da UFAM, da UNICAMP, bem como cooperações internacionais, com o professor Mohammed Zaki do Departamento de Ciência da Computação da RPI[13].

[13] `www.cs.rpi.edu\~zaki`

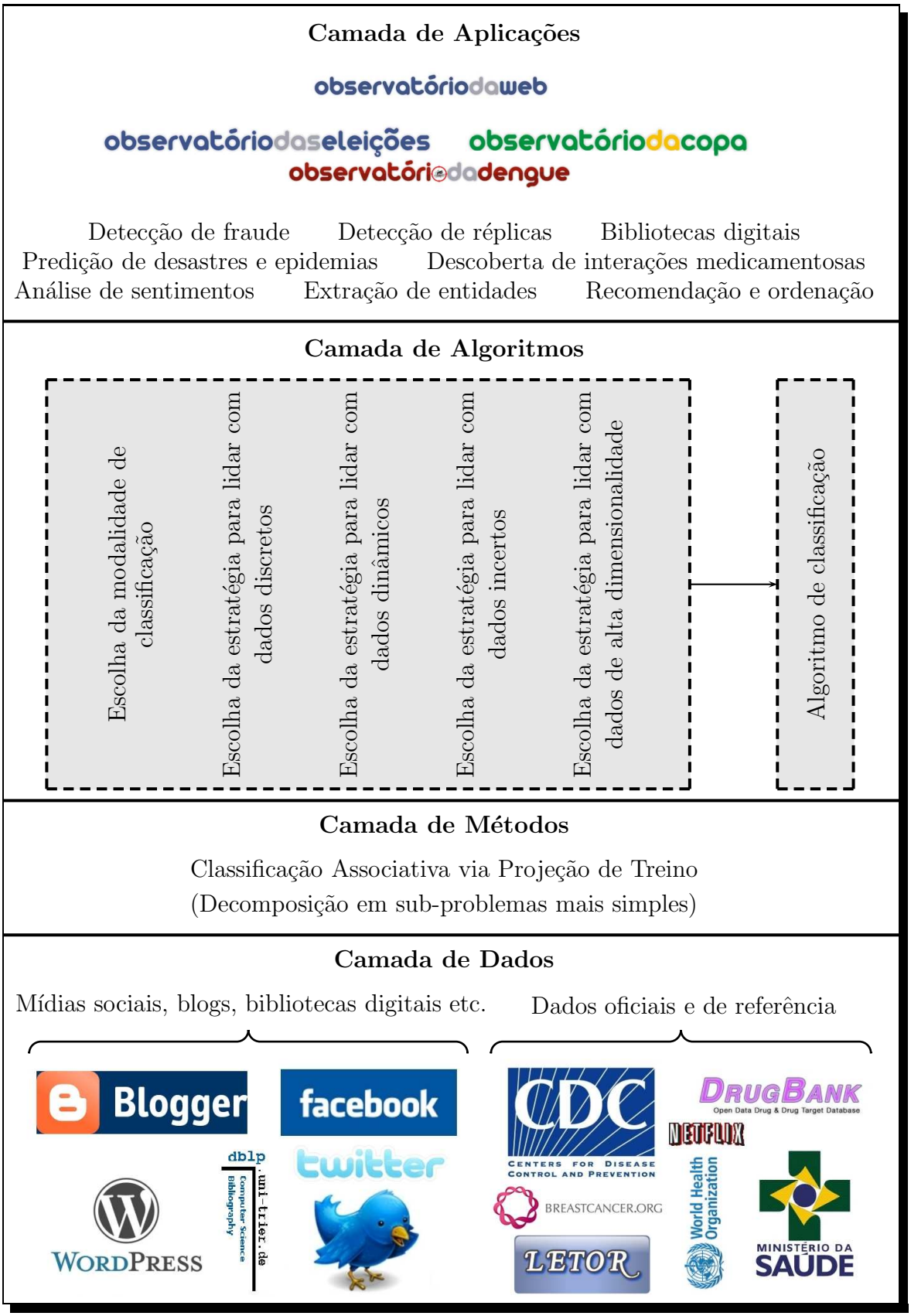

Figura 6: Estrutura em camadas do projeto.

4 Resultados

Nesta seção discutimos os resultados obtidos nos últimos 3 anos, bem como os resultados esperados ao fim deste projeto.

4.1 Resultados obtidos nos últimos 3 anos

Nos últimos três anos tivemos bons resultados de pesquisa abordando temas relacionados ao projeto proposto:

- Publicamos 8 artigos em periódicos, e 21 artigos em conferências, 12 das quais internacionais. Temos ainda 4 submissões para periódicos em avaliação. Ressaltamos a publicação de um livro em 2011 pela Springer.
- Formamos 2 alunos de iniciação científica. Formamos 2 mestres e mais 2 mestres devem ser formados até o final de 2011.
- Recebemos 6 premiações/distinções (5 nacionais e 1 internacional) e participamos de 9 projetos de pesquisa (4 deles como coordenador).
- Participamos de comitês de programa (eventos nacionais e internacionais) e revisamos artigos para pelo menos uma dezena de periódicos.

Uma relação detalhada de nossos indicadores de pesquisa é apresentada no Anexo A.

4.2 Resultados Esperados

Ao fim deste projeto (36 meses) espera-se obter os seguintes resultados:

- Projetar, implementar e validar novos métodos e algoritmos para classificação via projeção de treino.
- Avaliar a efetividade prática dos algoritmos desenvolvidos em aplicações relevantes e desafiadoras.
- Formar 2 doutores, 8 mestres e 10 alunos de iniciação científica.
- Publicar 6 artigos em periódicos, 8 artigos em conferências internacionais e 8 artigos em conferências nacionais.
- Publicar e transferir toda a tecnologia produzida durante o projeto para fins de pesquisa e desenvolvimento tecnológico.

5 Recursos

Nesta seção discutimos a demanda e disponibilidade de recursos necessários para a execução do projeto proposto.

5.1 Bolsa de Produtividade

Este projeto tem por objetivo principal a obtenção da bolsa de produtividade em pesquisa do proponente, a qual é um pilar fundamental para a execução do projeto.

5.2 Recursos de Pessoal

Os demais recursos de pessoal para a realização do projeto estão disponíveis. Os alunos que trabalham nas linhas de pesquisa já estão cursando doutorado, mestrado ou atuando como bolsistas de iniciação científica. Acreditamos que eventuais substituições não afetarão significativamente o trabalho.

5.3 Recursos de Equipamento

Em termos de equipamentos, acreditamos que estejamos em condições de suprir as demandas de desenvolvimento e avaliação inerentes ao projeto. Além da infra-estrutura do laboratório SPEED, recentemente renovado e estendido com recursos de projetos do CNPq e Fapemig. Mais ainda, o proponente é membro do Instituto Nacional de Ciência e Tecnologia para a Web (INWeb)[14], sediado no DCC-UFMG.

Referências

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD*, pages 207–216, 1993.

[2] **A. Veloso**, H. Almeida, M. Gonçalves, and W. Meira Jr. Learning to rank at query-time using association rules. In *ACM SIGIR*, pages 267–274, 2008.

[3] **A. Veloso**, B. Gusm ao, M. de Carvalho, and W. Meira Jr. Real world association rule mining. In *BNCOD*, pages 77–89, 2002.

[4] **A. Veloso**, B. Gusm ao, W. Meira Jr., M. de Carvalho, S. Parthasarathy, and M. Zaki. Efficiently mining approximate models of associations in evolving databases. In *PKDD*, pages 435–448, 2002.

[5] **A. Veloso**, B. Gusm ao, W. Meira Jr., and M. Bunte de Carvalho. Geração eficiente de regras de associação em bases de dados dinâmicas. In *CSBC*, pages 81–88, 2002.

[6] **A. Veloso**, A. Ferreira, M. Gonçalves, A. Laender, and W. Meira Jr. Cost-effective on-demand associative author name disambiguation. *Information Processing and Management*, publicação prevista para 2011.

[7] **A. Veloso**, M. Gonçalves, and W. Meira Jr. Competence-conscious associative rank aggregation. *Journal of Information and Data Management*, publicação prevista para 2011.

[14] `http://inctweb.dcc.ufmg.br`

[8] **A. Veloso**, M. Gonçalves, W. Meira Jr., and H. Almeida. Learning to rank using query-level rules. *Journal of Information and Data Management*, 1(3):567–582, 2010.

[9] **A. Veloso** and W. Meira Jr. Rule generation and rule selection techniques for cost-sensitive associative classification. In *SBBD*, pages 295–309, 2005.

[10] **A. Veloso** and W. Meira Jr. Lazy associative classification for content-based spam detection. In *LA-WEB*, pages 154–161, 2006.

[11] **A. Veloso** and W. Meira Jr. Efficient on-demand opinion mining. In *SBBD*, pages 332–346, 2007.

[12] **A. Veloso** and W. Meira Jr. Demand-driven associative classification. In *CSBC*, pages 10–18, 2010.

[13] **A. Veloso** and W. Meira Jr. *Demand-Driven Associative Classification*. Springer, 2011.

[14] **A. Veloso**, W. Meira Jr., M. Cristo, M. Gonçalves, and M. Zaki. Multi-evidence, multi-criteria, lazy associative document classification. In *ACM CIKM*, pages 218–227, 2006.

[15] **A. Veloso**, W. Meira Jr., M. de Carvalho, B. Pôssas, S. Parthasarathy, and M. Zaki. Mining frequent itemsets in evolving databases. In *SDM*, 2002.

[16] **A. Veloso**, W. Meira Jr., and M. Bunte de Carvalho. Mining reliable models of associations in dynamic databases. In *SBBD*, pages 263–277, 2002.

[17] **A. Veloso**, W. Meira Jr., and M. Bunte de Carvalho. Efficient data mining for frequent itemsets in dynamic and distributed database. In *CSBC*, pages 10–16, 2004.

[18] **A. Veloso**, W. Meira Jr., R. Ferreira, D. Guedes, and S. Parthasarathy. Asynchronous and anticipatory filter-stream based parallel algorithm for frequent itemset mining. In *PKDD*, pages 422–433, 2004.

[19] **A. Veloso**, W. Meira Jr., M. Gonçalves, H. Almeida, and M. Zaki. Calibrated lazy associative classification. *Information Sciences*, 181(13):2656–2670, 2011.

[20] **A. Veloso**, W. Meira Jr., and S. Parthasarathy. New parallel algorithms for frequent itemset mining in very large databases. In *SBAC-PAD*, pages 158–166, 2003.

[21] **A. Veloso**, W. Meira Jr., S. Parthasarathy, and M. Bunte de Carvalho. Efficient, accurate and privacy-preserving data mining for frequent itemsets in distributed databases. In *SBBD*, pages 281–292, 2003.

[22] **A. Veloso**, W. Meira Jr., and M. Zaki. Lazy associative classification. In *IEEE ICDM*, pages 645–654, 2006.

[23] **A. Veloso**, W. Meira Jr., and M. Zaki. Calibrated lazy associative classification. In *SBBD*, pages 135–149, 2008.

[24] **A. Veloso**, M. Otey, S. Parthasarathy, and W. Meira Jr. Parallel and distributed frequent itemset mining on dynamic datasets. In *HiPC*, pages 184–193, 2003.

[25] **A. Veloso**, B. Pôssas, G. Siqueira, W. Meira Jr., and M. de Carvalho. Mineração incremental de regras de associação. In *SBBD*, 2001.

[26] **A. Veloso**, M. Zaki, W. Meira Jr., and M. Gonçalves. Competence-conscious associative classification. *Statistical Analysis and Data Mining*, 2(5-6):361–377, 2009.

[27] **A. Veloso**, M. Zaki, W. Meira Jr., and M. Gonçalves. The metric dilemma: Competence-conscious associative classification. In *SIAM SDM*, pages 918–929, 2009.

[28] F. Benevenuto, T. Rodrigues, **A. Veloso**, J. Almeida, M. Gonçalves, and V. Almeida. Practical detection of spammers and content promoters in video sharing systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, publicação prevista para 2011.

[29] L. Bouarfa, P. Jonker, and J. Dankelman. Discovery of high-level tasks in the operating room. *Journal of Biomedical Informatics*, 44(3):455–462, 2011.

[30] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth Intl., 1984.

[31] P. Calais, **A. Veloso**, W. Meira Jr., and V. Almeida. From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In *ACM SIGKDD*, pages 81–90, 2011.

[32] P. Calais, T. Porto, L. Cerf, **A. Veloso**, and W. Meira Jr. V. Almeida. Exploiting temporal locality to determine user bias in microblogging platforms. *Journal of Information and Data Management*, publicação prevista para 2011.

[33] A. Davis, F. Peixoto, W. Santos, W. Meira Jr., **A. Veloso**, A. Soares, and A. Laender. Rt-ned: Real-time named entity disambiguation on twitter streams. In *SBBD-DEMO*, 2011.

[34] A. Davis, W. Santos, **A. Veloso**, A. Soares, A. Laender, and W. Meira Jr. Semi-supervised named entity disambiguation in streaming data. In *IEEE ICDE*, 2012 (submetido).

[35] J. de Freitas, G. Pappa, A. Soares, M. Gonçalves, E. de Moura, **A. Veloso**, A. Laender, and M. de Carvalho. Active learning genetic programming for record deduplication. In *IEEE CEC*, pages 1–8, 2010.

[36] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[37] F. Famili and S. Létourneau. Monitoring of aircraft operation using statistics and machine learning. In *ICTAI*, pages 279–286, 1999.

[38] F. Faria, **A. Veloso**, H. Almeida, E. Valle, R. Torres, M. Gonçalves, and W. Meira Jr. Learning to rank for content-based image retrieval. In *SIGMM Multimedia Information Retrieval*, pages 285–294, 2010.

[39] F. Faria, R. Calumby, **A. Veloso**, A. Rocha, and R. Torres. Uso de técnicas de aprendizado de máquina para classificação e recuperação de imagens. In *Workshop de Teses e Dissertações - SIBGRAPI*, 2011.

[40] K. Fenner, J. Gao, S. Kramer, L. Ellis, and L. Wackett. Data-driven extraction of relative reasoning rules to limit combinatorial explosion in biodegradation pathway prediction. *Bioinformatics*, 24:2079–2085, 2008.

[41] A. Ferreira, **A. Veloso**, M. Gonçalves, and A. Laender. Effective self-training author name disambiguation in scholarly digital libraries. In *ACM/IEEE JCDL*, pages 39–48, 2010.

[42] A. Ferreira, M. Gonçalves, J. Almeida, A. Laender, and **A. Veloso**. Generating synthetic authorship records for evaluating name disambiguation methods in scholarly digital libraries. *Information Sciences*, publicação prevista para 2012.

[43] A. Ferreira, M. Gonçalves, J. Almeida, A. Laender, and **A. Veloso**. Sygar - a synthetic data generator for evaluating name disambiguation methods. In *ECDL*, pages 437–441, 2009.

[44] M. Ganapathibhotla and B. Liu. Mining opinions in comparative sentences. In *COLING*, pages 241–248, 2008.

[45] J. Gomide, **A. Veloso**, W. Meira Jr., V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *ACM WebSci*, pages 1–8, 2011.

[46] B. Gu, B. Liu, F. Hu, and H. Liu. Efficiently determine the starting sample size for progressive sampling. In *Data Mining and Knowledge Discovery*, 2001.

[47] L. Hyafil and R. Rivest. Constructing optimal binary decision trees is np-complete. *Information Processing Letters*, 5(1):15–17, 1976.

[48] U. Kang, D. Chau, and C. Faloutsos. Mining large graphs: Algorithms, inference, and discoveries. In *IEEE ICDE*, pages 243–254, 2011.

[49] U. Kang, C. Tsourakakis, A. Appel, C. Faloutsos, and J. Leskovec. Hadi: Mining radii of large graphs. *TKDD*, 5(2):8, 2011.

[50] X. Li and B. Liu. Learning to classify texts using positive and unlabeled data. In *IJCAI*, pages 587–594, 2003.

[51] D. Lieb, A. Lookingbill, and S. Thrun. Adaptive road following using self-supervised learning and reverse optical flow. In *Robotics: Science and Systems*, pages 273–280, 2005.

[52] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *ACM SIGKDD*, pages 80–86, 1998.

[53] R. Marimont and M. Shapiro. Nearest neighbour searches and the curse of dimensionality. *Journal of the Institute of Mathematics and its Applications*, 24:59–70, 1979.

[54] G. Menezes, J. Almeida, F. Belém, M. Gonçalves, A. Lacerda, E. Moura, G. Pappa, **A. Veloso**, and N. Ziviani. Demand-driven tag recommendation. In *ECML/PKDD*, pages 402–417, 2010.

[55] R. Milasi, M. Jamali, and C. Lucas. Intelligent washing machine: A bioinspired and multi-objective approach. *International Journal of Control, Automation, and Systems*, 5(4):436–443, 2007.

[56] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[57] M. Otey, S. Parthasarathy, C. Wang, **A. Veloso**, and W. Meira Jr. Parallel and distributed methods for incremental frequent itemset mining. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(6):2439–2450, 2004.

[58] M. Otey, C. Wang, S. Parthasarathy, **A. Veloso**, and W. Meira Jr. Mining frequent itemsets in distributed and dynamic databases. In *IEEE ICDM*, pages 617–620, 2003.

[59] L. Pitt and L. Valiant. Computational limitations on learning from examples. *J. ACM*, 35(4):965–984, 1988.

[60] M. Ribeiro, **A. Veloso**, W. Meira Jr., L. Teixeira, P. Calais, and D. Guedes. Detecção de spams utilizando conteúdo web associado a mensagens. In *SBRC*, 2011.

[61] M. Ribeiro, P. Calais, **A. Veloso**, and W. Meira Jr. and. Spam detection using web page content: a new battleground. In *ACM CEAS*, pages 10–19, 2011.

[62] M. Ribeiro, W. Meira Jr., D. Guedes, and **A. Veloso**. Detecção de spams utilizando conteúdo web associado a mensagens. In *CSBC*, pages 55–61, 2011.

[63] M. Ribeiro, A. Veloso, W. Meira Jr., G. Pappa, L. Cherchiglia, and G. Brunoro. Mining twitter for feelings and opinions. In *SBBD-DEMO*, 2010.

[64] I. Santana, G. Barbosa, **A. Veloso**, W. Meira Jr., and R. Ferreira. Análise adaptativa de fluxo de sentimento baseada em janela deslizante ativa. In *SBBD*, volume publicação prevista para 2011.

[65] I. Santana, J. Gomide, **A. Veloso**, W. Meira Jr., and R. Ferreira. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *ACM SIGIR*, pages 475–484, 2011.

[66] I. Santana, J. Gomide, G. Barbosa, W. Santos, W. Meira Jr., **A. Veloso**, and R. Ferreira. Observatório da dengue: Surveillance based on twitter sentiment stream analysis. In *SBBD-DEMO*, 2011.

[67] W. Santos, G. Pappa, W. Meira Jr., D. Guedes, A. Veloso, V. Almeida, A. Pereira, P. Calais, A. Silva, F. Mour ao, T. Magalh aes, L. Cherchiglia, and G. Brunoro. Observatório da web: Uma plataforma de monitoração, síntese e visualização de eventos massivos em tempo real. In *SEMISH*, pages 56–67, 2010.

[68] A. Schneider, H. Feussner, N. Navab, H. Lemke, P. Jonker, and J. Dankelman. Prediction of intraoperative complexity from preoperative patient data for laparoscopic cholecystectomy. *Artificial Intelligence in Medicine*, 52(3):169–176, 2011.

[69] R. Silva, M. Gonçalves, and **A. Veloso**. Rule-based active sampling for learning to rank. In *ECML/PKDD*, pages 221–236, 2011.

[70] R. Smith, B. Dike, B. Ravichandran, A. El-Fallah, and R. Mehra. The fighter aircraft lcs: A case of different lcs goals and techniques. In *Learning Classifier Systems*, pages 283–300, 1999.

[71] **A. Veloso**, W. Meira, T. Macambira, D. Guedes, and H. Almeida. Automatic moderation of comments in a large on-line journalistic environment. In *ACM ICWSM*, pages 234–237, 2007.

[72] S. Thrun. A personal account of the development of stanley, the robot that won the darpa grand challenge. *AI Magazine*, 27(4):69–82, 2006.

[73] A. Tuzhilin and B. Liu. Querying multiple sets of discovered rules. In *ACM SIGKDD*, pages 52–60, 2002.

[74] L. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[75] L. Valiant. Pragmatic aspects of complexity theory (panel). In *IFIP Congress*, pages 7–8, 1986.

A Sumário de Indicadores

Apresentamos na Tabela 3 o sumário de indicadores de produtividade do proponente nos últimos 3 anos e nos últimos 10 anos. Informações mais detalhadas e atualizadas podem ser encontradas no currículo Lattes do proponente[15].

Indicador	**3 anos**	**10 anos**
Publicações	31	60
H-Index (*Publish or Perish*)	–	13
Livros	1	1
Periódicos	8	9
Artigos em conferências	22	50
Orientações concluídas	4	4
Mestrado (co-orientação)	2	2
Iniciação Científica	2	2
Orientações em andamento	11	11
Doutorado (co-orientação)	2	2
Mestrado (co-orientação)	4	4
Mestrado (orientação)	1	1
Iniciação Científica	4	4
Prêmios e distinções científicas	6	10
Nacional	9	9
Internacional	1	1
Projetos de pesquisa na própria instituição	4	9
Coordenação	3	3
Participação	1	5
Projetos de pesquisa multi-institucional	2	5
Participação	2	5
Projetos de pesquisa de cooperação internacional	1	2
Participação	1	2
Projetos de pesquisa de org. públicas ou privadas	2	5
Coordenação	1	2
Participação	1	3
Revisor de periódicos	10	14
Nacionais	2	2
Internacionais	8	12
Eventos	2	2
TPC de conferências nacionais	1	1
TPC de conferências internacionais	1	1

Tabela 3: Sumários dos indicadores de pesquisa.

[15] `http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4762232P7`

A seguir apresentamos uma análise qualitativa dos indicadores de pesquisa nos últimos 3 anos.

Publicações

A seguir destacamos as publicações realizadas em forma de livros, periódicos conferências internacionais e nacionais. Os artigos podem ser obtidos na íntegra em `www.dcc.ufmg.br/~adrianov`.

Livros

A. Veloso, W. Meira Jr. *Demand-Driven Associative Classification*. Springer, 2011, ISBN 978-0-85729-525-5, 125 páginas.

Periódicos

1. F. Benevenuto, T. Rodrigues, A. Veloso, J. Almeida, M. Gonçalves, and V. Almeida. Practical detection of spammers and content promoters in video sharing systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, publicação prevista para 2011 (QUALIS-A1).

2. A. Ferreira, M. Gonçalves, J. Almeida, A. Laender, A. Veloso. Generating synthetic authorship records for evaluating name disambiguation methods in scholarly digital libraries. *Information Sciences*, publicação prevista para 2011 (QUALIS-A1).

3. A. Veloso, A. Ferreira, M. Gonçalves, A. Laender, W. Meira Jr. Cost-effective on-demand associative author name disambiguation. *Information Processing and Management*, publicação prevista para 2011 (QUALIS-A2).

4. A. Veloso, M. Gonçalves, W. Meira Jr. Competence-conscious associative rank aggregation. *Journal of Information and Data Management*, publicação prevista para 2011.

5. P. Calais, T. Porto, L. Cerf, A. Veloso, W. Meira Jr. V. Almeida. Exploiting temporal locality to determine user bias in microblogging platforms. *Journal of Information and Data Management*, publicação prevista para 2011 ▶ participação de aluno de doutorado (P. Calais).

6. A. Veloso, W. Meira Jr., M. Gonçalves, H. Almeida, M. Zaki. Calibrated lazy associative classification. *Information Sciences*, 181(13):2656–2670, 2011 (QUALIS-A1) ▶ participação de aluno de doutorado (H. Almeida).

7. A. Veloso, M. Gonçalves, W. Meira Jr., H. Almeida. Learning to rank using query-level rules. *Journal of Information and Data Management*, 1(3):567–582, 2010 ▶ participação de aluno de doutorado (H. Almeida).

8. A. Veloso, M. Zaki, W. Meira Jr., M. Gonçalves. Competence-conscious associative classification. *Statistical Analysis and Data Mining*, 2(5-6):361–377, 2009.

Conferências Internacionais

1. P. Calais, A. Veloso, W. Meira Jr., V. Almeida. From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In *ACM SIGKDD*, pages 81–90, 2011 (QUALIS-B1) ▶ participação de aluno de doutorado (P. Calais).

2. I. Santana, J. Gomide, A. Veloso, W. Meira Jr., R. Ferreira. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *ACM SIGIR*, pages 475–484, 2011 (QUALIS-A1) ▶ participação de aluno de mestrado (I. Santana).

3. J. Gomide, A. Veloso, W. Meira Jr., V. Almeida, F. Benevenuto, F. Ferraz, M. Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *ACM WebSci*, pages 1–8, 2011.

4. R. Silva, M. Gonçalves, A. Veloso. Rule-based active sampling for learning to rank. In *ECML/PKDD*, pages 221–236, 2011 (QUALIS-B1) ▶ participação de aluno de mestrado (R. Silva).

5. M. Ribeiro, P. Calais, A. Veloso, and W. Meira Jr. and. Spam detection using web page content: a new battleground. In *ACM CEAS*, pages 10–19, 2011 ▶ participação de aluno de mestrado e de aluno de doutorado (M. Ribeiro e P. Calais).

6. J. de Freitas, G. Pappa, A. Soares, M. Gonçalves, E. de Moura, A. Veloso, A. Laender, M. de Carvalho. Active learning genetic programming for record deduplication. In *IEEE CEC*, pages 1–8, 2010. (QUALIS-A1)

7. G. Menezes, J. Almeida, F. Belém, M. Gonçalves, A. Lacerda, E. Moura, G. Pappa, A. Veloso, N. Ziviani. Demand-driven tag recommendation. In *ECML/PKDD*, pages 402–417, 2010 (QUALIS-B1) ▶ participação de aluno de mestrado (G. Menezes).

8. F. Faria, A. Veloso, H. Almeida, E. Valle, R. Torres, M. Gonçalves, W. Meira Jr. Learning to rank for content-based image retrieval. In *SIGMM Multimedia Information Retrieval*, pages 285–294, 2010 ▶ participação de aluno de doutorado (H. Almeida).

9. A. Ferreira, A. Veloso, M. Gonçalves, A. Laender. Effective self-training author name disambiguation in scholarly digital libraries. In *ACM/IEEE JCDL*, pages 39–48, 2010.

10. A. Veloso, M. Zaki, W. Meira Jr., M. Gonçalves. The metric dilemma: Competence-conscious associative classification. In *SIAM SDM*, pages 918–929, 2009.

11. A. Ferreira, M. Gonçalves, J. Almeida, A. Laender, A. Veloso. Sygar - a synthetic data generator for evaluating name disambiguation methods. In *ECDL*, pages 437–441, 2009 (QUALIS-B1).

12. A. Veloso, H. Almeida, M. Gonçalves, W. Meira Jr. Learning to rank at query-time using association rules. In *ACM SIGIR*, pages 267–274, 2008 (QUALIS-A1) ▶ participação de aluno de doutorado (H. Almeida).

Conferências Nacionais

1. A. Veloso, W. Meira Jr. Demand-driven associative classification. In *CSBC*, pages 10–18, 2010.

2. A. Veloso, W. Meira Jr., M. Zaki. Calibrated lazy associative classification. In *SBBD*, pages 135–149, 2008.

3. A. Davis, W. Santos, W. Meira Jr., A. Veloso, A. Soares, A. Laender. Rt-ned: Real-time named entity disambiguation on twitter streams. In *SBBD-DEMO*, 2011 ▶ participação de aluno de iniciação científica (A. Davis).

4. F. Faria, R. Calumby, A. Veloso, A. Rocha, R. Torres. Uso de técnicas de aprendizado de máquina para classificação e recuperação de imagens. In *Workshop de Teses e Dissertações - SIBGRAPI*, 2011.

5. M. Ribeiro, A. Veloso, W. Meira Jr., L. Teixeira, P. Calais, D. Guedes. Detecção de spams utilizando conteúdo web associado a mensagens. In *SBRC*, 2011 ▶ participação do aluno de iniciação científica e de aluno de doutorado (M. Ribeiro e P. Calais).

6. M. Ribeiro, W. Meira Jr., D. Guedes, A. Veloso. Detecção de spams utilizando conteúdo web associado a mensagens. In *CSBC*, pages 55–61, 2011 ▶ participação de aluno de iniciação científica (M. Ribeiro).

7. I. Santana, G. Barbosa, A. Veloso, W. Meira Jr., R. Ferreira. Análise adaptativa de fluxo de sentimento baseada em janela deslizante ativa. In *SBBD*, volume publicação prevista para 2011 ▶ participação de aluno de mestrado (I. Santana).

8. I. Santana, J. Gomide, G. Barbosa, W. Santos, W. Meira Jr., A. Veloso, R. Ferreira. Observatório da dengue: Surveillance based on twitter sentiment stream analysis. In *SBBD-DEMO*, 2011 ▶ participação de aluno de mestrado (I. Santana).

9. M. Ribeiro, A. Veloso, W. Meira Jr., G. Pappa, L. Cherchiglia, G. Brunoro. Mining twitter for feelings and opinions. In *SBBD-DEMO*, 2010 ▶ participação de aluno de iniciação científica (M. Ribeiro).

10. W. Santos, G. Pappa, W. Meira Jr., D. Guedes, A. Veloso, V. Almeida, A. Pereira, P. Calais, A. Silva, F. Mour ao, T. Magalh aes, L. Cherchiglia, G. Brunoro. Observatório da web: Uma plataforma de monitoração, síntese e visualização de eventos massivos em tempo real. In *SEMISH*, pages 56–67, 2010 ▶ participação de aluno de doutorado (P. Calais).

Orientações

A seguir destacamos as orientações defendidas e em andamento.

Tese de Doutorado – Atualmente estamos co-orientando dois alunos de doutorado. Ambas as orientações estão em andamento:

Co-orientação de Pedro Calais Guerra (P. Calais), Desenvolvimento de Algoritmos de Transferência de Aprendizado aplicados a Mídias Sociais, defesa prevista para 2013. Artigos produzidos [60, 31, 61, 32, 67].

Co-orientação de Humberto Mossri de Almeida (H. Almeida), Desenvolvimento de Algoritmos de Transferência de Aprendizado aplicados a Mídias Sociais, defesa prevista para 2013. Artigos produzidos [8, 19, 38, 2].

Dissertação de Mestrado – Dois alunos defenderam suas dissertações sob a co-orientação do proponente, e mais 5 estão atualmente em andamento:

Co-orientação de Guilherme Vale Menezes (G. Menezes), Recomendação de *Tags* Sob-Demanda, defendida em 2011. Artigos produzidos [54].

Co-orientação de Rickson Guidolini (R. Guidolini), Detecção de Réplicas de Sítios Web em Máquinas de Busca usando Aprendizado de Máquina, defendida em 2011.

Orientação de Gessé Dafé (G. Dafé), Classificação Associativa através de Combinação por Proximidade de Atributos, defesa prevista para 2013.

Co-orientação de Aline Bessa (A. Bessa), Detecção de Sítios Replicados usando Aprendizado Ativo, defesa prevista para 2013.

Co-orientação de Marco Túlio Ribeiro (M. Ribeiro), Recomendação de Produtos através de Agregação de Listas Ordenadas, defesa prevista para 2012.

Co-orientação de Ismael Santana (I. Santana), Análise Adaptativa de Fluxo de Sentimento baseada em Janela Deslizante Ativa, defesa prevista para 2011. Artigos produzidos [65, 66, 64].

Co-orientação de Rodrigo Silva (R. Silva), Amostragem Ativa aplicada á Ordenação de Documentos Web, defesa prevista para 2011. Artigos produzidos [69].

Iniciação Científica – Dois alunos concluíram seus trabalhos, e mais 4 alunos estão atualmente sob a orientação do proponente:

Alexandre Guelman Davis (A. Davis), Bolsa BITIB de iniciação científica, concluída em 2011.

Marco Túlio Ribeiro (M. Ribeiro), Bolsa CNPQ de iniciação científica, concluída em 2011. Artigos produzidos [62, 60, 61, 63].

Alexandre Guelman Davis (A. Davis), UOL Bolsa Pesquisa de iniciação científica, término em 2012. Artigos produzidos [33, 34]

Felipe Peixoto (F. Peixoto), Bolsa PIBIC de iniciação científica, término em 2011. Artigos produzidos [34].

Luis Matoso (L. Matoso), Bolsa CNPQ de iniciação científica, término em 2012.

André Harder (A. Harder), Bolsa CNPQ de iniciação científica, término em 2012.

Palestras Convidadas

Fomos convidados para proferir duas palestras:

- Yahoo! Research Barcelona, Demand-Driven Associative Classification, Barcelona, 20 de setembro de 2010.
- UNICAMP, Classificação Associativa Sob-Demanda, Campinas, 10 de junho de 2010.

Projetos de Pesquisa

Atualmente participamos e coordenados 9 projetos de pesquisa:

1. Pesquisador associado ao Instituto Nacionalde Ciência e Tecnologia para a Web. Processo MCT/CNPQ573871/2008-6. Valor: R$2.300.000,00. Início em 2010, término em 2012.
2. Membro da linha de pesquisa Descoberta de Conhecimento em Bases de Dados *i* InWeb. Valor: R$278.048,57.
3. Coordenador do projeto CNPQ-UNIVERSAL, "Algoritmos de Aprendizado Associativo pata Ordenação de Documentos". Processo MCT/CNPQ483829/2010-2. Valor R$20.000,00. Início em 2011, término em 2013.
4. Coordenador do projeto FIAT-FAPEMIG, "Análise Espaço-Temporal de Opiniões acerca de Modelos Automotivos". Processo APQ-03829-10. Valor R$9.072,00. Início em 2011, término em 2012.
5. Coordenador do projeto BITIB-FAPEMIG, "Análise em Tempo Real de Sentimentos na Web". Valor R$6.000,00. Início em 2010, término em 2011.
6. Coordenador do projeto UOL-Bolsa Pesquisa, "Classificação Associativa Sob-Demanda em Tempo Real e perante Dados com Incerteza". Processo 20110215172500. Valor R$22.000,00. Início em 2011, término em 2012.
7. Participante do projeto CNPQ-PDI, "Modelos e Algoritmos para Tratamento de Informações em Tempo Real". Processo MCT/CNPQ/560286/2010-4. Valor R$299.786,32. Início em 2011, término em 2012.
8. Participante do projeto UOL-PAGSEGURO, "Detecção de Fraudes Eletrônicas". Início em 2011, término em 2012.
9. Participante do projeto SERPRO-DATA MINING, "Mineração de Dados em Larga Escala". Início em 2011, término em 2012.

Prêmios e Distinções

A seguir listamos as premiações e distinções recebidas:

1. Primeiro lugar no Concurso de Iniciação Científica (CTIC) - 2002 (com o artigo [5]), promovido pela Sociedade Brasileira de Computação.
2. Primeiro lugar no Concurso de Teses e Dissertações (CTD) - 2004 (com o artigo [17]), promovido pela Sociedade Brasileira de Computação.
3. Primeiro lugar no Concurso de Teses e Dissertações (CTD) - 2010 (com o artigo [12]), promovido pela Sociedade Brasileira de Computação.
4. Prêmio UFMG de teses - 2010 (`http://www.ufmg.br/online/arquivos/017552.shtml`).
5. Grande Prêmio UFMG de teses (menção honrosa) - 2010.
6. Melhor artigo do Simpósio Brasileiro de Banco de Dados - 2002 (com o artigo [16]).
7. Entre os 3 melhores artigos do Simpósio Brasileiro de Banco de Dados - 2003 (com o artigo [21]).
8. Entre os 3 melhores artigos da SIAM Data Mining Conference - 2009 (com o artigo [27]).
9. Entre os 5 melhores artigos do Simpósio Brasileiro de Redes de Computadores - 2011 (com o artigo [60]).
10. Primeiro lugar no Concurso de Iniciação Científica (CTIC) - 2011, promovido pela Sociedade Brasileira de Computação (trabalho do aluno Marco Túlio Ribeiro [62]).

Participação em Comitês de Programa e Revisão de Periódicos

Participamos de comitês de programa de conferências, e revisamos artigos para diversos periódicos:

1. Revisor de artigos para IEEE Transactions on Parallel and Distributed Systems
2. Revisor de artigos para Data Mining and Knowledge Discovery
3. Revisor de artigos para Information Systems
4. Revisor de artigos para Journal of the Brazilian Computer Society
5. Revisor de artigos para IEEE Transactions on Knowledge and Data Engineering
6. Revisor de artigos para Journal of Data and Information Management
7. Revisor de artigos para Journal of Autonomous Agents and Multi-Agent Systems
8. Revisor de artigos para ACM Transactions on Intelligent Systems and Technology

9. Revisor de artigos para The VLDB Journal
10. Membro do comitê de programa do ACM WWW Conference 2012, Lion, França.
11. Membro do comitê de programa do SBBD 2011, Florianólis, Brasil.