O Uso de UMLS para Aprimorar a Recomendação de Graus de Vigilância para Pacientes do Setor Primário

Flávia Pena Nicolas[1,2], Juliana Tarossi Pollettini[1], Sylvia G. Panico[2], Julio C. Daneluzzi[2], Alessandra Alaniz Macedo[1]

[1] Grupo de Informática Biomédica
Departamento de Física e Matemática, FFCLRP-USP

[2] Faculdade de Medicina de Ribeirão Preto, FMRP-USP

{flavianicolas, jupollettini, sylviapanico}@gmail.com,
juliodaneluzzi@hotmail.com, ale.alaniz@usp.br

Abstract. *A rating index called "Grau de Vigilância – GV" (Surveillance Level) supports the identification of people requiring different levels of medical care. This paper investigate the use of Unified Medical Language System (UMLS) aiming to optimize results from machine learning algorithms when suggesting patient's GV. We run experiments with patients attended by "Centro Médico Social e Comunitário de Vila Lobato" (CMSC Vila Lobato), Ribeirão Preto/SP.*

Resumo. *O índice de classificação para atendimento, Grau de Vigilância – GV, apóia a identificação de indivíduos que requerem diferentes níveis de atendimento médico. O presente artigo propõe investigar o uso do Sistema Unificado de Linguagem Médica (UMLS), na tentativa de otimizar resultados de classificação de paciente com algoritmos de Aprendizado de Máquina na determinação automática do GV de pacientes do Centro Médico Social e Comunitário de Vila Lobato (CMSC Vila Lobato) em Ribeirão Preto/SP.*

1. Introdução

As informações coletadas na área da atenção básica podem auxiliar na prevenção de doenças e aprimorar a qualidade de vida dos pacientes. Tecnologias de software e de hardware podem aprimorar a manipulação e o intercâmbio da informação coletada. Para intercâmbio de informações, mapeamentos entre vocabulários biomédicos tornam-se fundamentais.

Para ajudar na identificação de indivíduos que requerem diferentes graus de atendimento, o Centro Médico de Vila Lobato (CMSC Vila Lobato) utiliza uma medida de classificação para atendimento denominada Grau de Vigilância (GV). O GV é estabelecido segundo a existência e associação a fatores de risco à saúde e ao desenvolvimento aos quais estão expostos os pacientes e suas famílias e varia em valores gradativos possibilitando a definição de ações educativas, terapêuticas ou especializadas requeridas pelos pacientes e respectivas famílias [2].

O sistema computacional GV-Automático propôs a definição automática ou semi-automática, em alguns casos especiais, do GV do paciente a partir da manipulação e análise de informações provenientes dos prontuários médicos [1]. O GV-Automático possui módulos de classificação automática de informações a partir de conceitos de Vizinhos Próximos, Redes Neurais Artificiais, *Relevance Feedback*, Árvores de Decisão e a combinação desses conceitos [1, 2, 3, 4]. Um dos problemas enfrentados na criação desses módulos era a falta de padronização dos termos usados nos registros médicos. Após análise de dados de atendimentos, percebeu-se que, por exemplo, condutas prescritas utilizavam

"Orientação alimentar", "Orientações alimentares" e "Orientação quanto à alimentação" para expressar recomendações idênticas. Além das variações sintáticas, encontram-se variações semânticas significativas no vocabulário médico como cosmecêutico, dermatocosmético, cosmético funcional, bioativo e neocêutico. Assim, acreditava-se que os resultados obtidos pelo GV-Automático estivessem sendo influenciados pelas diferenças linguísticas da linguagem de especialidade médica e de saúde e que artefatos lingüísticos pudessem auxiliar na determinação automática do GV.

Este artigo propõe investigar os benefícios que artefatos lingüísticos como os de ontologias e dicionários do Sistema Unificado de Linguagem Médica (*Unified Medical Language System* – UMLS) [5] podem trazer à pesquisa de categorização de pacientes por GV. A partir do estudo aprofundado de UMLS e de padronizações manuais e eletrônicas, as informações existentes nos registros médicos como IDs (Impressões Diagnósticas) e CDs (Condutas Médicas) foram padronizadas de maneira automática e agrupadas em categorias para serem usadas como atributos/valores para os algoritmos de Aprendizado de Máquina (AM). Visava-se aperfeiçoar os resultados da determinação automática do GV do paciente.

O presente artigo encontra-se organizado da seguinte forma: a Seção 2 resume os fundamentos teóricos estudados, a Seção 3 descreve a proposta com as atividades realizadas, a Seção 4 discute experimentos iniciais e a Seção 5 apresenta a conclusão.

2. Fundamentos Teóricos

Para o desenvolvimento da proposta foi realizado o estudo dos fundamentos teóricos relacionados aos temas da investigação aqui apresentada.

2.1. Sistema Unificado de Linguagem Médica

O Sistema Unificado de Linguagem Médica (UMLS) foi desenvolvido pela *National Library of Medicine* (NLM) para ajudar profissionais da saúde e pesquisadores a acessar informações biomédicas de várias origens [6]. O projeto buscou criar mecanismos que facilitassem a circulação de informações. Hoje sua versão mais recente é a 2009AB [7].

O UMLS é uma coleção de referências cruzadas de dados médicos e uma série de fontes de conhecimento, que pretende auxiliar a troca de informações no domínio da Saúde, atacando o problema da multiplicidade de sistemas de códigos existentes atualmente [5]. Ele é composto por 3 fontes de conhecimento: *Metathesaurus*, um grande depósito, com mais de 1 milhão de conceitos biomédicos de mais de 100 origens; *Semantic Network*, uma rede limitada de 135 tipos semânticos e 54 relações entre as categorias; e o *SPECIALIST Lexicon*, com informação lexical e programas para o processamento de linguagem. Essas fontes podem ser usadas juntas ou separadamente.

Enquanto a estrutura de cada origem é preservada na construção do *Metathesaurus*, termos equivalentes são agrupados em um conceito semanticamente único. Relações interconceituais são também herdadas de vocabulários implícitos (básicos) ou geradas especificamente. Como o *Metathesaurus* impôs que não houvesse restrições nas fontes (origens), não é possível fornecer o tipo de organização esperada de uma ontologia. Em contraste, o *Semantic Network* é desenvolvido independente de vocabulários integrados no *Metathesaurus* e serve como uma básica ontologia de alto-nível para o domínio biomédico [7]. Assim, tipos semânticos são usados para categorizar todos os conceitos UMLS [8].

No nível mais alto, o *Semantic Network* é organizado em torno de oposições de entidades e eventos, e duas hierarquias de herança única refletem esta distinção. Os filhos imediatos de *Entidade* são *Objeto Físico* e *Entidade Conceitual*, enquanto *Evento* tem *Atividade* e *Fenômeno* ou *Processo* como descendentes diretos. Cada tipo semântico na rede

tem uma definição textual e aparece em uma dessas hierarquias. Em adição à taxonomia, relações associativas em cinco subcategorias são definidas entre os tipos semânticos: Física (por exemplo: *part_of, branch_of, ingredient_of*), Espacial (por exemplo: *location_of, adjacent_to*), Funcional (por exemplo: *treats, complicates, causes*), Temporal (por exemplo: *co-occurs_with, precedes*) e Conceitual (*evaluation_of, diagnosis*). Desde que cada conceito do *Metathesaurus* é atribuído a pelo menos um tipo semântico, relações entre tipos semânticos também definem a semântica permitida para as relações entre os conceitos [9]. O UMLS pode ser usado, entre outras coisas, na recuperação de informação, construção de thesaurus, processamento de linguagem natural, indexação automatizada e registros eletrônicos de saúde (EHR) [5]. Ao final do estudo sobre UMLS, planejou-se investigar o *Metathesaurus* para suporte da proposta.

2.2 Ontologias

Uma ontologia é uma especificação de conceitualização. Ela descreve os conceitos e relações que podem existir e formaliza a terminologia em um domínio [10]. Ontologias são usadas para facilitar o compartilhamento de conhecimento entre as pessoas, processamento de informação, mineração de dados, comunicação entre agentes de software, ou outras aplicações em processamento de conhecimento [11].

O propósito de uma ontologia é estudar classes de entidades (por exemplo: substâncias, qualidades e processos) na realidade em que determinada ontologia têm significância biomédica. Diferentemente de terminologias biomédicas, que coletam os nomes das entidades empregadas no domínio biomédico, ontologias biomédicas estão preocupadas com o princípio da definição das classes biológicas e relações entre elas. Na prática, como elas são mais do que listas de termos, mas não necessariamente encontram os requisitos de uma organização formal, os muitos produtos desenvolvidos por terminologistas e ontologistas biomédicos sempre caem entre terminologias e ontologias, e constituem um “gradiente ontológico” [13].

Muitas ontologias têm sido desenvolvidas no campo biomédico. O UMLS, apoiado pela NLM, é o principal meio para facilitar programas de computador a processar e gerenciar documentos biomédicos [12]. As fontes de conhecimento do UMLS têm sido amplamente usadas em Processamento de Linguagem Natural (PLN).

3. Trabalho Realizado

As investigações de aprimoramento da classificação de pacientes de acordo com GVs por meio do uso de artefatos lingüísticos do UMLS foram realizadas em etapas conforme apresentação das subseções a seguir. Os resultados das tarefas realizadas foram abstraídos em algoritmos.

3.1. Elaboração de arquivos de padronização manual e eletrônica

Após investigações e leituras do domínio do trabalho, foi realizada a elaboração manual de dois arquivos de padronização: um de impressões diagnósticas (IDs) e outro de condutas médicas (CDs). Como fonte de conhecimento utilizou-se registros médicos armazenados em tabelas do banco de dados do CMSC Vila Lobato. Em seguida, um arquivo de texto com o conteúdo de cada tabela foi gerado, para que pudesse ser feita a análise e posterior padronização dos termos semelhantes (ver Algoritmo 1).

A partir da leitura dos arquivos, cada termo (CD ou ID) foi analisado individualmente e, em seguida, foi realizada uma busca (manual) por termos sinônimos ou semelhantes no mesmo arquivo. Se o termo foi encontrado, ele é adicionado à linha em que

estava seu semelhante. Para as IDs com seus sinônimos foram atribuídos valores do Código Internacional de Doenças – versão 10 (CID-10). Por exemplo, “Adenomegalia cervical” e “Adenomegalia reacional” possuem o mesmo CID-10, R59.0. Já a “Adenopatia inguinal” possui o CID-10 com valor R59.9. Neste trabalho, esses dois conceitos foram considerados sinônimos, uma vez que ambos pertencem ao mesmo capítulo.

Após esse tratamento textual, houve uma redução no tamanho dos arquivos (21% para as CDs e 37,5% para as IDs). Essas reduções podem diminuir o tempo de processamento das informações por sistemas de informação em saúde. Porém o principal foco era prover sistematizações e padronizações de informações médicas para auxiliar a gestão hospitalar e a obtenção de conhecimento médico.

Elaboração Manual de Arquivo de Padronização:
Passo 1: Gerar um arquivo de IDs e um de CDs a partir das informações contidas no BD do CMSC Vila Lobato
Passo 2: ***Para cada*** *arquivo*
Para cada *termo*
Realizar uma busca manual por termos sinônimos ou semelhantes
Se *o termo é encontrado: adicioná-lo para linha em que está seu semelhante*
Passo 3: ***Para*** *o arquivo de IDs*
Considerar também o CID-10 no momento de buscar pelo(s) termo(s) sinônimo(s)

Algoritmo 1. Abstração do mapeamento manual de Impressões Diagnósticas (IDs) e Condutas Médicas (CDs).

Para efeito de comparação foi criado um arquivo de padronização de maneira eletrônica, mas dessa vez apenas para IDs, justificando-se pelo fato de que as CDs são termos e expressões generalistas, i.e., não restritas ao domínio médico. Uma vez que as IDs são expressões presentes no CID-10, isso as torna universais, sendo possível a realização de uma busca com resultados consideráveis.

A ferramenta usada para o mapeamento eletrônico foi a MetamorphoSys [15], na versão 2007AC. MetamorphoSys é a ferramenta assistente e de customização para instalação UMLS. Essa ferramenta instala fontes de conhecimento UMLS. Se o Metathesaurus for selecionado, a criação customizada de subconjuntos Metathesaurus torna-se possível. O MetamorphoSys pode ser usado para excluir vocabulários que não são necessários ou licenças para uso em aplicações locais e para seleção a partir de uma variedade de opções de dados de saída e filtros. O Metathesaurus consiste em vários arquivos com diferentes funcionalidades e com mais de 100 vocabulários descritos. Algumas fontes requerem licenças separadas para uso específico dessa fonte de conhecimento. Assim, o Metathesaurus torna-se muito útil para consultas e pesquisas de termos em diferentes vocabulários biomédicos.

Para a montagem do arquivo de padronização, cada ID do arquivo original foi buscada no MetamorphoSys com suporte do Metathesaurus e, se encontrada, foi adicionada ao arquivo (ver Algoritmo 2). Ao final, percebeu-se que houve uma redução no arquivo de padronização eletrônica de quase 7% em relação ao arquivo original. Este fato pode ser explicado por terem sido usados termos e fontes em português, uma vez que este idioma ainda não está muito difundido no UMLS. Se fosse utilizado o inglês, por exemplo, provavelmente o resultado apresentado seria mais abrangente.

Como não houve a criação de um arquivo de padronização eletrônica para CDs, apenas os arquivos de padronização das IDs foram comparados. Como esperado, a

padronização manual apresentou melhores resultados. Essa espera deve-se ao fato de que o agente humano, que procura um a um, cada termo, embute no resultado o seu conhecimento com relação às definições e contextos correspondentes. Esse tipo de conhecimento é muito difícil de ser atingido usando um software como o MetamorphoSys. Além disso, na padronização manual foi utilizado também o valor do CID-10 como fator para definir se as expressões eram sinônimas ou não.

Elaboração Eletrônica de Arquivo de Padronização:
Passo 1: Gerar um arquivo de IDs a partir das informações contidas no BD do CMSC Vila Lobato
Passo 2: ***Para cada*** *termo*
Buscar o termo no MetamorphoSys (ferramenta do UMLS)
Se *o termo for encontrado*
Analisar os termos sinônimos que foram retornados
Se *o sinônimo constar no BD do CMSC Vila Lobato*
Adicionar o termo para a linha em que está seu semelhante

Algoritmo 2. Abstração do mapeamento eletrônico de Impressões Diagnósticas (IDs).

3.2. Processamento textual para padronização automática de IDs/CDs

Para importar os termos e conceitos contidos no UMLS utilizou-se um sistema de gerenciamento de banco de dados (SGBD) que ofereceu a possibilidade de, durante sua instalação e configuração, serem selecionadas opções de acordo com as necessidades do usuário, como utilizar apenas os vocabulários de interesse. No caso deste projeto desejava-se utilizar apenas termos em português, por isso foram instaladas as quatro fontes neste idioma presentes no UMLS: MedDRA (*Medical Dictionary for Regulatory Activities Terminology*), WHOPOR , MSHPOR (tradução para portugês do *Medical Subject Headings* - MeSH) e ICPC (*The International Classification of Primary Care*) [16].

Na codificação do algoritmo de padronização automática foram úteis as colunas CUI (Identificador Único de Conceito) do Metathesaurus, que atribui um código para cada conceito, ou seja, termos diferentes e com o mesmo conceito possuem o mesmo valor do CUI, e STR, referente a string em si. Foram então criadas três classes em Java: *ConectarAoBanco*, *Comparar* e o *Main*. A classe *ConectarAoBanco* é composta pelo construtor *ConectarAoBanco()* e pelos métodos *ConectarAoPostgre(), ConectarAoMSAccess(), getResultSet(String sql), executarSQL(String sql)* e *fecharConexao().* Esta classe, com os métodos citados anteriormente, é a responsável por estabelecer a conexão com o PostgreSQL, onde estão contidos os dados necessários do banco de dados do CMSC Vila Lobato (Condutas e Impressões Diagnósticas) e estabelecer a conexão com o SGBD, onde estão contidos os termos do UMLS que foram comparados com as IDs e CDs do CMSC Vila Lobato. A classe *ConectarAoBanco* executa e armazena os resultados de uma busca SQL que foi passada como parâmetro e, por fim, fecha a conexão. Já a classe *Comparar* possui os métodos *CompararCondutas()* e *CompararIDs(),* além do construtor *Comparar().* Está é a classe que realiza a comparação em si. O método *CompararCondutas()* e o *CompararIDs()* funcionam da mesma maneira: a conexão com o PostgreSQL é estabelecida e logo após uma consulta é realizada (seleção das CDs ou das IDs), sendo o resultado desta consulta armazenado em uma lista *impressaod* ou *cdta*. Em seguida, a conexão com o SGBD é estabelecida e 2 consultas são feitas: uma selecionando os termos (String da coluna STR) e outra os Identificadores Únicos de Conceito (coluna CUI). Estes são armazenados em duas listas diferentes: *str_umls* e *cui_umls*.

Quando todos os dados necessários já foram armazenados, cada termo original das listas de CDs ou IDs é comparado a cada termo da lista de termos do UMLS (*str_umls*). Caso um termo igual seja encontrado, o CUI deste é armazenado e então é realizada uma busca na lista de CUIs (cui_umls), visando encontrar CUIs com o mesmo valor deste, o que significa que os conceitos são iguais. Quando CUIs de mesmo valor são encontrados, os termos referentes a eles são adicionados a uma nova lista (ArrayList), que armazena o resultado final das buscas e comparações. Nessa lista, os termos considerados sinônimos são agrupados e separados por vírgula e os diferentes grupos são separados por uma barra ("/"). Se o termo buscado não existe na lista de termos do UMLS, ele não aparece na lista dos termos agrupados. Quando todas as comparações foram realizadas e os termos agrupados, o método *fecharConexao()* é chamado para encerrar a conexão com o banco.

A classe *Main()* chama os métodos *CompararCondutas()* e *CompararIDs()* e quando executada exibe uma lista de padronização automática de Condutas e uma de Impressões Diagnósticas. O processamento descrito é apresentado no Algoritmo 3 e sua esquematização de acordo com a codificação da proposta é resumida na Figura 1.

Elaboração Automática de Arquivo de Padronização:
Passo 1: Importar os termos e conceitos em português contidos no UMLS através de um SGBD com funcionalidade específica de escolha de termos
Passo 2: Estabelecer a conexão com o PostgreSQL (onde estão contidos os dados do CMSC Vila Lobato) e realizar consulta ao banco armazenando as CDs e IDs em duas listas diferentes
Passo 3: : Estabelecer a conexão com o SGBD (onde estão contidos os termos e conceitos do UMLS) e realizar consulta ao banco armazenando o termo (STR) e o Identificador Único de Conceito (CUI)
Passo 4: ***Para cada*** *lista de CDs e para cada lista de IDs*
    Para cada *termo*
        Comparar este com cada termo na lista do UMLS
        Se *um termo igual é encontrado*
            Armazenar o CUI referente ao termo
            Buscar por CUIs de mesmo valor, com a finalidade de encontrar conceitos sinônimos
            Se *CUIs iguais são encontrados*
                Armazenar os termos referentes e eles em uma nova lista com o resultado das buscas e comparações

Algoritmo 3. Abstração da geração automática do arquivo de padronização.

3.3. Agrupamento das IDs/CDs em categorias para serem utilizadas como atributos/valores para os algoritmos de Aprendizado de Máquina

Após a execução do algoritmo de padronização automática, foi criado a partir da lista de IDs gerada um arquivo para ser utilizado como atributo/valor para os algoritmos de AM. Para auxiliar na construção deste arquivo outro arquivo de IDs foi utilizado, montado em projeto anterior, baseado nos Capítulos e Valores do CID-10, além do arquivo original de IDs, gerado a partir de informações de IDs (Impressões Diagnósticas) contidas no banco de dados do CMSC Vila Lobato.

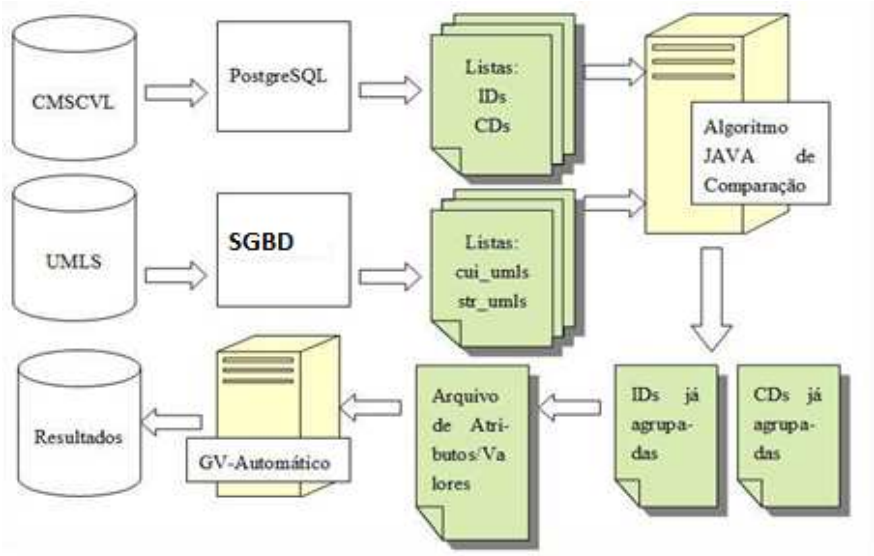

Figura 1. Processo de implementação do algoritmo de padronização automática, criação de arquivo de Atributos/Valores para ser usado na classificação automática de pacientes a partir do GV (GV-Automático) e obtenção dos resultados.

Na elaboração do arquivo, as IDs foram separadas de acordo com seu tipo, baseado nos capítulos do CID-10, sendo que os Atributos foram atribuídos à descrição do capítulo em si (os tipos) e os Valores Assumidos foram os próprios termos contidos no banco de dados do CMSC Vila Lobato, a partir dos quais foi realizada a busca por termos considerados sinônimos.

A partir da lista de IDs, os termos pertencentes ao grupo de cada Valor Assumido foram designados Sub-Valores, inclusive o próprio valor. Quando havia termos idênticos apenas um exemplar de cada termo era acrescentado ao arquivo.

A lista de Condutas não foi útil para a montagem de um arquivo de padronização para ser utilizado posteriormente como atributo/valor em algoritmos de AM, pois, de um total de 2.452 condutas médicas, apenas 28 foram encontradas no banco de dados do UMLS (1,15% do total). Das 28 condutas, 19 possuíam termos considerados sinônimos, mas, em todos os casos, esses termos sinônimos constavam apenas no banco de dados do UMLS e não do CMSC Vila Lobato, o que para o estudo atual não possui representatividade.

4. Experimentos

Os experimentos da proposta foram realizados a partir do sistema GV-Automático de [2]. Assim, o arquivo resultante do processamento descrito na Figura 1 foi utilizado como Atributo/Valor em algoritmos de AM do GV-Automático. O algoritmo avaliado foi o *K*-Vizinhos Mais Próximos (*k-Nearest Neighbor*) [17], uma vez que ele foi o primeiro algoritmo explorado no GV-Automático. O conceito de *k*-Vizinhos Mais Próximos fundamenta-se nessa aprendizagem baseada em instâncias e, segundo Russell e Norvig, sua idéia-chave consiste no fato de que as propriedades de qualquer ponto de entrada específico têm probabilidade de serem semelhantes às propriedades de outros pontos em sua vizinhança [18]. Nesse sentido, o vizinho mais próximo de um padrão de entrada *x* é definido como sendo o que apresenta a menor distância entre seu vetor de características e o vetor de características de *x*. Ao classificar um novo padrão de entrada *x*, esse modelo se baseia numa função de distância para determinar o quão próximo o novo exemplo se encontra dos padrões armazenados e utiliza a classe associada ao vizinho mais próximo para determinar a classe do novo exemplo.

Inicialmente foi atribuído o valor 1 para *K* e então selecionou-se o período dos atendimentos a serem analisados, que foi de 01/01/2001 até 01/11/2009, utilizando o modo de treinamento e teste *cross-validation* com 10 *folds*. Nesse caso a porcentagem de acertos foi de 38,57%, ou seja, 206/534. Alterando-se o valor de vizinhos para 2, o total de acertos foi 37,45% (200/534). Quando o valor de vizinhos foi alterado para 3, obteve-se um melhor resultado, com 39,32% de acertos (210/534). Foram ainda realizados testes para 4, 5 e 6 vizinhos, e as taxas de acertos obtidas foram, respectivamente, 39,70% (212/534), 40,44% (216/534) e 40,82% (218/534), conforme Figura 2. Como foi possível perceber, com exceção de 2 vizinhos, à medida que *K* aumenta o número de acertos também aumenta. Isso, na verdade, não é uma regra, pois depois de um determinado valor a tendência é que a precisão caia. A seleção do valor de *K* é de grande influência para os resultados: se *K* é muito alto, alguns desses vizinhos podem ter probabilidades diferentes; se *K* é muito baixo, a estimativa pode não ter credibilidade.

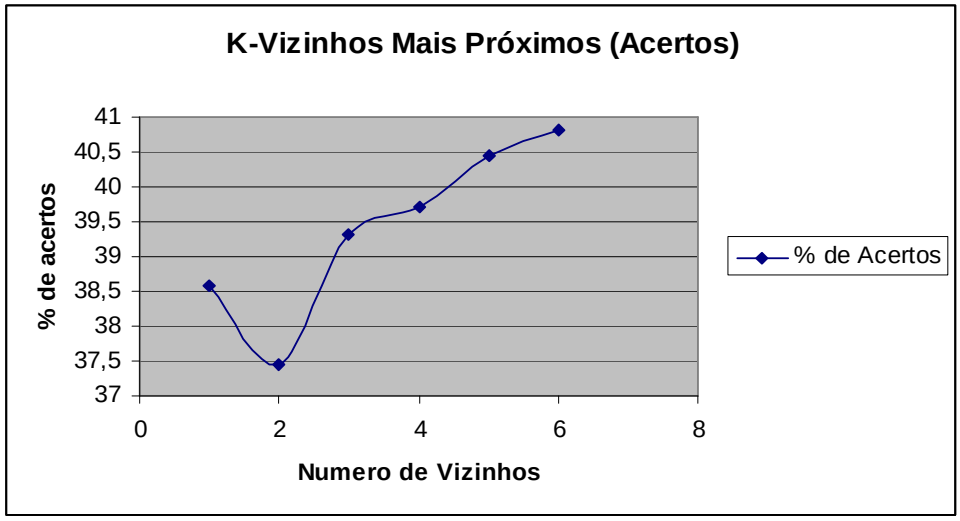

Figura 2. Porcentagem de acertos do *K*-Vizinhos Mais Próximos

Comparando-se esses resultados com os obtidos em estudo anterior, pode-se perceber, de acordo com a Figura 3, uma pequena e aparente melhora, uma vez que para 1 vizinho o resultado anterior foi 37,25% de acertos, enquanto este foi de 38,57% de acertos. Esta situação pode ser explicada pelo fato de que, mesmo encontrando muitos sinônimos no banco de termos do UMLS, como mostrado no arquivo de padronização, estes não foram suficientes, uma vez que não eram os termos usados pelos profissionais do CMSC Vila Lobato, não constando em seu Banco de Dados.

5. Conclusão

O algoritmo de padronização automática foi criado com sucesso, possibilitando o agrupamento de CDs e IDs de acordo com seu significado, utilizando o UMLS como ferramenta essencial. A partir do agrupamento gerado, criou-se um arquivo para ser utilizado como Atributo/Valor no algoritmo de AM *K*-Vizinhos Mais Próximos do sistema de GV-Automático [1,2,3,4].

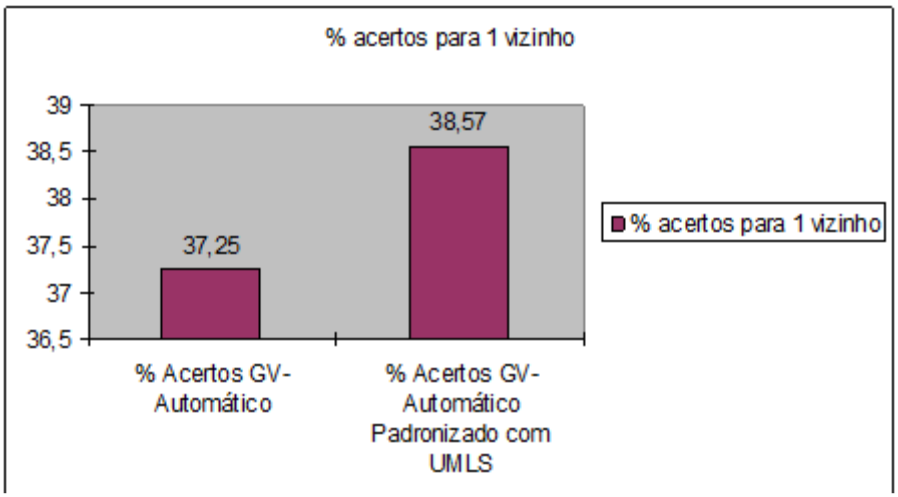

Figura 3. Porcentagem de acertos para 1 vizinho, sem e com padronização de termos

A partir dos resultados obtidos nos testes de software observou-se uma aparente melhora na determinação automática do GV de pacientes em relação ao estudo anterior, em que não houve padronização dos termos considerados sinônimos. Esta suposta melhora inicialmente não foi significativa, o que pode ser justificado pelo fato de que, mesmo encontrando muitos sinônimos no banco de termos do UMLS estes não foram suficientes, uma vez que não eram os termos usados pelos profissionais do CMSC Vila Lobato. Testes estatísticos estão sendo realizados a fim de comprovar o aprimoramento dos resultados com o uso de UMLS.

Para trabalhos futuros pretende-se realizar testes de software utilizando Árvores de Decisão, buscando um melhor desempenho em relação ao algoritmo *K*-Vizinhos Mais Próximos, além de estudar a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) a fim de aplicá-la ao problema em questão. Esta é uma metodologia para desenvolvimento de sistemas de suporte a decisão e consiste em um conjunto de fases e processos padrões, não rígidos e independentes da área de negócio e das ferramentas utilizadas, de forma estruturada e metódica.

6. Agradecimentos

Os autores agradecem a Pró-Reitoria de Pesquisa da Universidade de São Paulo (RUSP) e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

7. Referências

[1] Pollettini, J. T.; Tinos, R.; Panico, S.; Daneluzzi, J. C.; Macedo, A. A., Classificação automática de pacientes para atendimento médico pediátrico multidisciplinar a partir do seu Grau de Vigilância. In: Workshop de Informática Médica. Anais do XXVIII Congresso da Sociedade Brasileira de Computação, Belém-Pará-Brasil, p. 61-70, 2008.

[2] Pollettini, J. T.; Nicolas, F. P.; Panico, S.; Daneluzzi, J. C.; Tinos, R.; Baranauskas, J. A.; Macedo, A. A.. A software architecture-based framework supporting suggestion of medical surveillance level from classification of electronic patient records. In: The 12th IEEE International Conference on Computational Science and Engineering, Vancouver-Canadá, 8p (electronically published), 2009.

[3] Pollettini, J. T.; Panico, S.; Daneluzzi, J. C.; Tinos, R.; Baranauskas, J. A.; Macedo, A. A.. Computer-aided decision for medical surveillance level from classification of electronic patient records. Submetido ao periódico International Journal of Medical Informatics.

[4] Pollettini, J. T.; Tinos, R.; Panico, S.; Daneluzzi, J. C.; Macedo, A. A.. Vigilância em atenção básica à saúde a partir do uso de relevance feedback para classificação de pacientes em diferentes níveis de cuidado em saúde. In: IX Workshop de Informática Médica. Anais do Congresso da Sociedade Brasileira de Computação, Bento Gonçalves - RS – Brasil, p. 1945-1954, 2009.

[5] UMLS Home Page. Disponível em: www.nlm.nih.gov/research/umls/pdf/UMLS_Basics.pdf

[6] Lindberg, D. A., Humphreys, B. L., and Mc Cray, A.T. The Unified Medical Language System, Methods Inf Med 32(4):281-291

[7]Metathesaurus. Disponível em:nlm.nih.gov/research/umls/licensedcontent/downloads.html

[8] McCray. An upper-level ontology for the biomedical domain". Comparative and Functional Genomics Comp Funct Genom 4: 80–84, 2003.

[9] McCrayA.; Nelson. The Representation of meaning in the UMLS. Methods Inf Med 34(1-2):193-201, 1995.

[10] McCray A.T.; Bodenreider O. A Conceptual Framework for the Biomedical Domain. The Semantics of Relationships: An Interdisciplinary Perspective. Boston: Kluwer Academic Publishers, 181-198, 2002.

[11] Grüninger, M.; Lee, J. Ontology Applications and Design -Introduction. Commun. ACM 45(2): 39-41, 2002.

[12] Medical Informatics – Knowledge Management and Data Mining in Biomedicine, edited by H. Chen, S.S. Fuller, C. Friedman, W. Hersh. Capítulo 1: H. Chen, S.S. Fuller, C. Friedman, W. Hersh.

[13] McCray et al., 1993; Humphreys et all.,1993; Campbell et al., 1998 Lindberg DA, Humphreys BL, McCray AT. 1993. The Unified Medical Language System. Methods Inf Med 32(4): 281–291.

[14] Medical Informatics – Knowledge Management and Data Mining in Biomedicine, edited by H. Chen, S.S. Fuller, C. Friedman, W. Hersh. Capítulo 8: O. Bodenreider, A. Burgun.

[15] Metamorphosys. Disponível: nlm.nih.gov/research/umls/about_umls.html#MetamorphoSys

[16] ICPC. Disponível em: www.nlm.nih.gov/research/umls/metaa1.htm

[17] Moreno-Seco, F.; Mico, L.; Oncina, J. A new classification rule based on nearest neighbour search. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. v.4, p.408-411, 2004.

[18] Russel, S. J.; Norvig, P. Métodos estatísticos de aprendizagem. Inteligência Artificial. Tradução: Vandenberg D. de Souza. 2.ed., p.690-737, 2004.