JHI JOURNAL OF HEALTH INFORMATICS

Pré-processamento para mineração de dados sobre beneficiários de planos de saúde suplementar

Data mining pre-processing for beneficiaries of health insurance

Procesamiento para la minerá de datos los beneficiarios de los planes de salud complementario

Everton Fernando Barros[1], Wesley Romão[2], Ademir Aparecido Constantino[3], Celso Lara de Souza[4]

RESUMO

Descritores: Base de Dados, Mineração de Dados, Planos de Pré-Pagamento em Saúde

Objetivo: O objetivo é preparar dados de um PSS para facilitar a utilização por algoritmos de mineração de dados (MD) e demonstrar uma metodologia para sua preparação. **Métodos:** Os métodos utilizados para preparar os dados foram propostos por Fayyad: entender o domínio da aplicação; criação de um conjunto de dados alvo; limpeza dos dados, redução e projeção dos dados. Essa metodologia foi aplicada de forma iterativa e interativa: iterativa porque realizou-se consultas a analistas de domínio e interativa porque alguns processos se repetem no decorrer da preparação. **Resultados:** Conseguiu-se organizar os dados, originalmente em um banco de dados relacional, em apenas uma tabela e reduzir o número de atributos em mais de 50%, além reduzir a quantidade de instâncias em 14%. **Conclusão:** Demonstrou-se um pré-processamento sobre dados de um PSS e obteve-se dados adequados para serem utilizados por algoritmos de MD.

ABSTRACT

Keywords: Database, Data Mining, Prepaid Health Plans

Objective: The goal is to prepare data from a HI to facilitate the use of data mining (DM) algorithms and demonstrate a methodology for preparation of such data. **Methods:** The methods used to prepare these data are the methods proposed by Fayyad: understanding the application domain, create a set of target data, data cleansing, data reduction and projection. This methodology are both iterative and interactive. It is iterative by interviews with the domain analysts. It is interactive because the processes are repeated during the preparation. **Results:** We managed to organize the data, originally in a relational database in just one table and reduce the number of attributes in more than 50%, and reduce the number of instances at 14%. **Conclusion:** It was shown a pre-processing data from a HI in which obtained data suitable for use by DM algorithms.

RESUMEN

Descriptores: Base de Datos, Minería de Datos, Planes de Salud de Prepago

Objetivo: El objetivo es preparar a los datos de un soporte técnico para facilitar el uso de algoritmos de minería de datos (MD) y demostrar una metodología para la preparación de estos datos. **Métodos:** Los métodos utilizados para la preparación de estos datos son los métodos propuestos por Fayyad: comprender el dominio de aplicación, crear un conjunto de datos objetivo, la depuración de datos, reducción de datos y de proyección. Esta metodología se aplicó en una consulta es decir, de manera iterativa con los analistas y el área interactiva en la que los procedimientos se repitieron durante la preparación. **Resultados:** Se ha podido organizar los datos, originalmente en una base de datos relacional en una sola mesa y reducir el número de atributos en más del 50%, y reducir el número de casos en 14%. **Conclusión:** Los resultados confirman un pre-tratamiento de los datos de un soporte técnico y obtener datos adecuados para el uso de algoritmos de DM.

[1] *Mestrando em Ciência da Computação pela UEM – Universidade Estadual de Maringá – UEM - Maringá (PR), Brasil.*
[2] *Professor Adjunto, Universidade Estadual de Maringá – UEM – Maringá (PR), Brasil.*
[3] *Professor Titular, Universidade Estadual de Maringá – UEM – Maringá (PR), Brasil.*
[4] *Pós-Graduação em Gestão Empresarial pela FGV, Diretor da Benner Tecnologia e Sistemas de Saúde Ltda, Maringá (PR), Brasil.*

Autor Correspondente: **Everton Fernando Barros**
e-mail: **efbarro@gmail.com**

Artigo recebido: **14/12/2010**
Aprovado: **01/03/2011**

INTRODUÇÃO

As operadoras de planos de saúde suplementar possuem uma grande quantidade de dados armazenados sobre procedimentos realizados pelos beneficiários. Essas bases de dados podem conter conhecimentos ocultos de alta qualidade que poderão auxiliar na tomada de decisões em programas de prevenção de doenças.

Para descoberta de conhecimento oculto em banco de dados é necessário a aplicação de algoritmos de mineração de dados (MD), parte do processo de descoberta de conhecimento em banco de dados (DCBD). No entanto, não importa o quão "inteligente" um algoritmo de MD seja, ele falhará na descoberta de conhecimento de alta qualidade se for aplicado em dados de baixa qualidade[1]. Como a maioria das bases de dados apresenta diversos problemas (como valores faltando, valores inconsistentes, falta de precisão, ruídos e erros de medição) que reduzem a sua qualidade, há necessidade de se aplicar técnicas de pré-processamento para melhoria da qualidade destes dados.

Considerando os diversos problemas normalmente encontrados em bases de dados aplicou-se técnicas de pré-processamento em dados de um plano de saúde suplementar (PSS) objetivando prepará-los para utilização em algoritmos de MD. No decorrer das etapas do pré-processamento alguns problemas foram encontrados, como por exemplo dados armazenados em banco de dados relacional. A maioria dos algoritmos de MD não conseguem trabalhar com dados armazenados dessa forma e para resolver este problema foi necessário analisar estes dados e identificar as tabelas e os atributos dessas tabelas para que pudessem ser utilizados no processo de DCBD.

Segundo autor[2], o DCBD consiste em nove etapas, mas o presente trabalho é fundamentado nas quatro primeiras: seleção de um conjunto de dados alvo, limpeza de dados, redução e projeção de dados.

Portanto, apresenta-se uma metodologia para aplicação destas etapas sobre dados de um PSS objetivando tanto demonstrar a metodologia quanto obter dados preparados para serem utilizados em algoritmos de MD desenvolvidos pelo grupo de pesquisa GPEA (Grupo de Pesquisa em Engenharia de Algoritmo) da UEM (Universidade Estadual de Maringá).

MÉTODOS

Os métodos utilizados neste trabalho baseiam-se nas quatro primeiras etapas do processo de DCBD proposto por Fayyad[2] com algumas adaptações. Os passos básicos dos métodos aplicados consistem em:

- Entendimento do domínio da aplicação que implica no levantamento de um conhecimento prévio relevante e dos objetivos do usuário;
- Seleção de um conjunto de dados alvo, ou seja, selecionar um conjunto de dados ou concentrar-se em um subconjunto ou instâncias sobre os quais a descoberta será efetuada;
- Limpeza de Dados que consiste em operações básicas tais como remoção de dados errôneos e manipulação de atributos com valores ausentes;
- Pré-seleção de atributos que são relevantes com base no domínio da aplicação e objetivos do usuário;
- Redução e Projeção de Dados que envolvem métodos de transformação para reduzir o número efetivo de atributos.

Esses métodos são aplicados e descritos mais detalhadamente nas próximas seções, levando em consideração o objetivo de preparar dados de um PSS.

Entendimento do Domínio da Aplicação

Apesar de o presente trabalho abordar apenas a preparação de um conjunto de dados para a aplicação de algoritmos de MD, o objetivo final desta pesquisa, considerando o processo completo de DCBD, é encontrar padrões válidos, compreensíveis e úteis para gestores de planos de saúde suplementar. Para encontrar esses padrões é necessário definir o escopo do processo de DCBD. Nesta pesquisa focou-se em procedimentos, especialidades e diagnósticos de doenças de beneficiários de um plano de saúde.

Os dados foram obtidos exclusivamente para fins de pesquisa, sem qualquer identificação pessoal, a partir de uma base de teste de uma operadora de saúde suplementar contendo informações administrativas e de procedimentos em hospitais, laboratórios e consultórios relativas a beneficiários de um plano de saúde do estado de Santa Catarina, disponibilizados na forma de um banco de dados em *SQL Server*.

É importante salientar que as informações pessoais, relativas aos beneficiários desse plano de saúde, foram excluídas da base preservando a privacidade dos beneficiários. Além disso, os resultados representam uma reorganização de dados já existentes não apresentando descrição de perfis ou sugestão de diagnósticos, excluindo a necessidade de aprovação do comitê de ética. Essa exclusão de informações pessoais não causa nenhum prejuízo ao processo de DCBD visto que o objetivo final da pesquisa é buscar padrões genéricos e não informação específica de uma pessoa.

Seleção de um Conjunto de Dados Alvo

Como se utilizou somente um Banco de Dados Relacional, a seleção do conjunto de dados alvo torna-se uma seleção de tabelas. Como o objetivo final da pesquisa é encontrar padrões sobre doenças, selecionaram-se algumas tabelas que tinham correlação com os beneficiários, tais como procedimentos médicos e diagnósticos, para extrair dados que possam satisfazer esse objetivo. Analisando essa base de dados e levando em consideração o objetivo de encontrar padrões sobre doenças, selecionaram-se as seguintes tabelas, GRS-BENEFICIARIO, GRS-GUIA, GRS-GUIA-EVENTO, SAM-TGE e SAM-CID cujas descrições seguem abaixo.

GRS-BENEFICIARIO - Todos os beneficiários inscritos no plano de saúde.

GRS-GUIA - Possui dados sobre os processos administrativos e de diagnósticos dos beneficiários.

GRS-GUIA-EVENTO - Possui eventos utilizados

pelos beneficiários.

SAM-TGE - É a tabela geral de eventos e possui a descrição de todos os eventos possíveis do plano.

SAM-CID - Essa tabela contém todos os códigos possíveis de CID (Classificação Internacional de Doenças) além de sua descrição.

Limpeza dos dados

No mundo real os dados geralmente são incompletos, inconsistentes e apresentam algum tipo de ruído. Limpeza de dados visa detectar e remover estas anomalias dos dados com o objetivo de aumentar/melhorar sua qualidade[3].

Na etapa de seleção do conjunto de dados alvo foi criada uma nova tabela com o histórico dos beneficiários na qual já foi realizada alguma limpeza nos dados, onde levou-se em consideração por exemplo se um beneficiário usou ou não uma especialidade ou teve ou não uma determinada doença. Isto eliminou problemas com valores nulos em muitos atributos pois considerou-se, por exemplo, se um paciente teve ou não determinada doença.

Segundo autor[3], para atingir o objetivo de descobrir conhecimento útil a partir de grandes bases de dados é importante considerar a possibilidade de retornar e avançar entre as fases do processo de DCBD repetindo certas etapas para obter um melhor resultado. Portanto, no decorrer do pré-processamento foram detectadas anomalias e ruídos nos dados, como por exemplo a detecção de idades com valores negativos, implicando que a etapa de limpeza dos dados fosse aplicada novamente.

Pré-seleção de atributos

Para se realizar uma seleção prévia dos atributos foi levado em consideração o fato de muitos atributos relevantes praticamente não possuírem dados, como o caso do atributo IMC (Índice de Massa Corpórea) da tabela GRS-BENEFICIÁRIO. Em seguida foram descartados mais alguns atributos que não continham valor atribuído a nenhum registro da tabela de beneficiários. A partir dos atributos remanescentes e das tabelas foi realizada a junção entre as tabelas e conseguiu-se retirar um histórico de procedimentos dos pacientes. Para fazer a junção encontrou-se vários problemas, como por exemplo cada beneficiário pode possuir várias guias e cada guia pode conter vários eventos. Isso faria, após o processo de junção, um beneficiário possuir vários registros o que impossibilitaria a idéia de histórico, ou seja, haveria vários registros de beneficiários para cada guia, e vários registros de guias para cada evento o que invalidaria a etapa de MD com muitos registros repetidos. Com base na junção preliminar das tabelas GRS-BENEFICIARIO, GRS-GUIA e GRS-GUIA-EVENTO, obteve-se uma tabela maior contendo os seguintes atributos.

BENEFICIARIO - Código identificador do beneficiário.

SEXO - Sexo do beneficiário.

IDADE - Idade do beneficiário.

ESTADO - Estado onde o beneficiário reside.

MUNICIPIO - Município onde o beneficiário reside.

TIPO-BENEF - Se o beneficiário é titular, ou dependente.

TIPO-GUIA - Se o beneficiário fez uma consulta, usou medicamentos, teve uma internação ou utilizou o SADT (Serviço Auxiliar de Diagnóstico e Terapia).

ESPECIALIDADE - Indica a especialidade utilizada pelo beneficiário para uma determinada GUIA.

EVENTO - Indica os eventos que o beneficiário utilizou-se para uma GUIA.

CID - Classificação Internacional de Doenças - identifica qual doença o beneficiário foi diagnosticado.

Como cada beneficiário pode utilizar muitas guias, várias especialidades, muitos códigos de CID e muitos eventos, fez-se necessário transformar os atributos TIPO-GUIA, ESPECIALIDADE, EVENTO e CID em vários outros atributos.

O atributo TIPO-GUIA possui como domínio quatro valores, SADT, CONSULTA, INTERNAÇÃO e MEDICAMENTOS. O valor MEDICAMENTOS não possui nenhum registro na base de dados, sendo, portanto descartado. A partir do atributo TIPO-GUIA obtiveram-se os atributos SADT, CONSULTA e INTERNAÇÃO no qual os valores possíveis é a quantidade de utilização de cada serviço por um beneficiário.

O atributo ESPECIALIDADE possui como domínio 69 especialidades. Então a partir desse atributo obteve-se 69 novos atributos que correspondem a cada especialidade disponibilizada pelo plano de saúde. O domínio de cada novo atributo contém a quantidade de vezes que um beneficiário utilizou a especialidade.

O atributo EVENTO possui como domínio 17324 eventos o que torna inviável a transformação desses valores em atributos. Entretanto as tabelas GRS-GRUPOEVENTOS e GRS-GRUPOEVENTOS-EVENTO agrupam eventos que se relacionam a alguma doença comum como por exemplo Diabetes Mellitus. Desses grupos criou-se então 11 atributos e, de acordo com os eventos realizados pelos beneficiários, incrementava-se o valor para o grupo no qual o evento pertencia. O domínio desses grupos é o número de vezes que cada beneficiário utilizou os eventos pertencentes a um determinado grupo de eventos.

O atributo CID possui como domínio 14202 códigos, inviabilizando a transformação desses valores em atributos. Entretanto encontrou-se na literatura[4] um agrupamento desses códigos de CID construído de forma empírica tendo como meta principal manter a coerência clínica dentro do mesmo capítulo da CID e um volume mínimo de internações em cada novo grupo para análise de re-internações hospitalares. Segundo os autores, apesar de ter foco em internação ou re-internações hospitalares, esses agrupamentos podem ser úteis para análise de problemas de saúde e também podem ser adaptados para outros objetivos. Utilizou-se esses agrupamentos para transformação desses dados em atributos totalizando 32 novos atributos, no qual o valor de cada atributo é o número de vezes que o paciente foi atendido com um CID pertencente a um determinado grupo de CID.

Redução e projeção de dados

Para se ter uma boa redução dos dados faz-se necessário conhecer como esses dados estão organizados, qual relevância de cada atributo e também qual a relação entre os atributos. Para ter esse conhecimento dos dados é necessário fazer algumas análises estatísticas. Nas próximas seções trataremos dessas análises.

Classificação Internacional de Doenças (CID)

Devido ao fato do atributo CID ser convertido em vários atributos, apresenta-se alguns gráficos para um melhor entendimento desses novos atributos gerados.

Analisando os grupos de códigos de CID formados percebeu-se que o agrupamento XXI (Contato com os serviços de saúde) concentra a maior parte do total de códigos, 65,72% das ocorrências ante 34,28% dos demais grupos. Esse grupo é formado pelos códigos que pertencem ao intervalo Z00 a Z99. Em uma consulta ao banco de dados verificou-se que mais de 98% desses registros pertencem a CID Z00 que, segundo a página na Internet do Ministério da Saúde[5], correspondem a exames em geral e investigações de pessoas sem queixas ou diagnósticos relatados. A maior parte desses códigos são casos no qual os médicos não informam o CID do paciente ou não houve diagnóstico.

Pela alta dimensionalidade do atributo GRUPO-CID-XXI em relação aos demais retirou-se o grupo XXI da geração do histograma de frequência de grupos de códigos de CID para uma melhor visualização e comparação.

Verifica-se, a partir do histograma da Figura 1, que há vários grupos de códigos de CID com baixa frequência o que pode dificultar ou inviabilizar a MD.

Como se preservou informações como por exemplo, a quantidade de vezes que um beneficiário foi internado ou a quantidade de vezes que foi tratar-se de uma determinada doença "CID", fez-se uma nova análise dos dados agora desconsiderando o número de vezes que o beneficiário foi atendido com uma determinada doença, considerando apenas se teve uma determinada doença ou não. Verificou-se então que 45,79% (114.487) das ocorrências são do grupo XXI e 54,21% (135.531) das ocorrências dos demais grupos. Portanto a proporção entre os demais grupos de códigos e o grupo XXI se inverte.

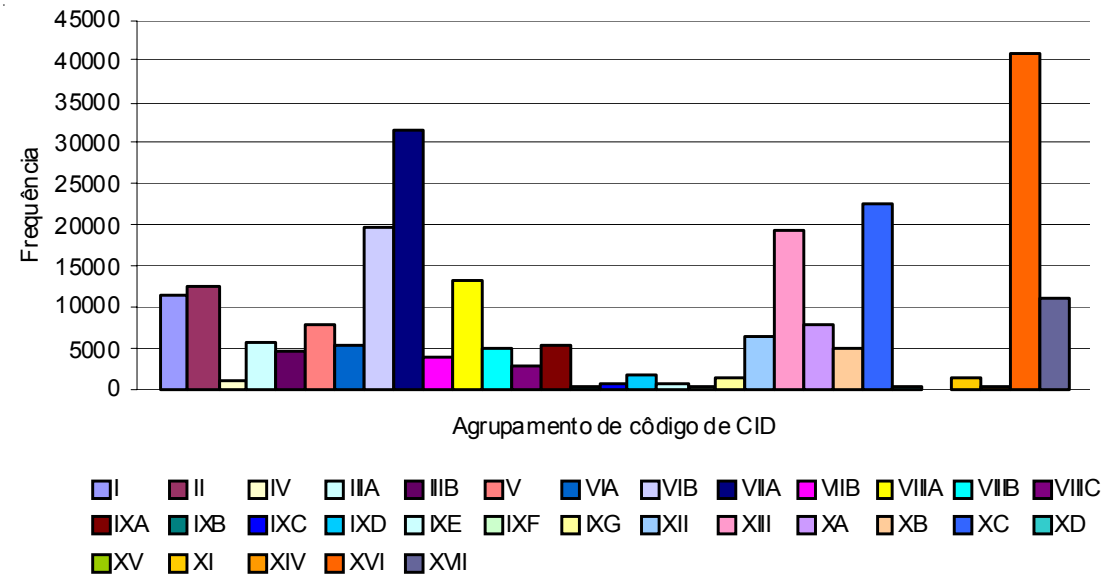

Figura 1 - Histograma frequência de agrupamentos de códigos de CID

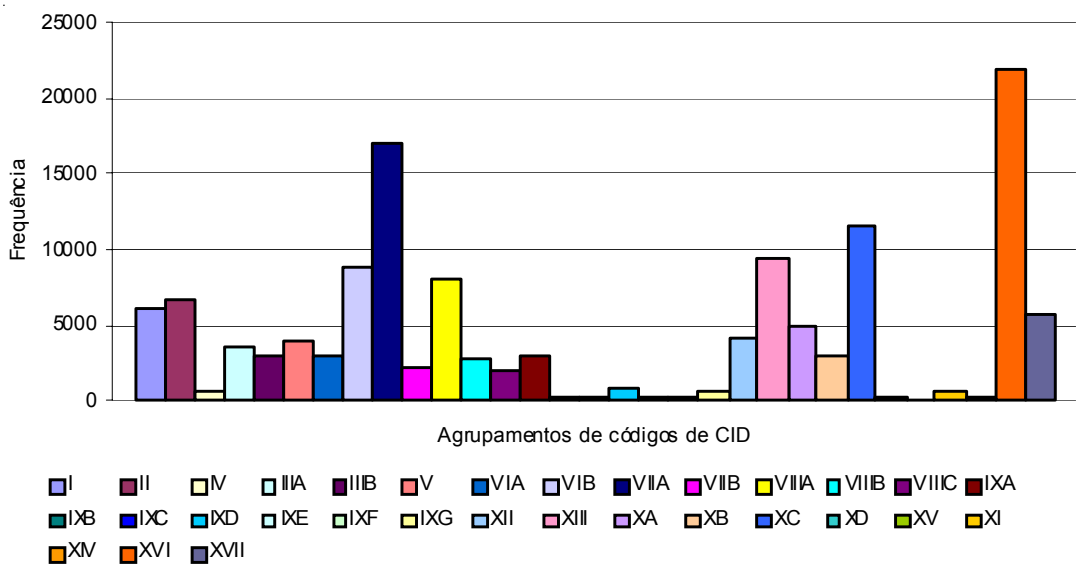

Figura 2 - Histograma de frequência de Figura 2- Agrupamentos de códigos de CID desconsiderando repetições

Mesmo havendo uma redução da proporção em relação ao total de registros o grupo XXI ainda possui uma alta dimensionalidade o que atrapalha a visualização de um histograma que compara a frequência de cada grupo de CID.

O histograma da Figura 2 leva em consideração se um beneficiário teve ou não uma determinada doença. Percebe-se que o comportamento do histograma da Figura 2 é muito similar ao histograma da Figura 1. Isso significa que os grupos com baixa frequência continuam com baixa frequência em relação aos de alta frequência o que pode ser um indicativo de quais grupos seriam melhor para aplicação dos algoritmos de MD.

Especialidades

Como o atributo ESPECIALIDADE foi convertido em 69 atributos fez-se necessário uma análise para um melhor entendimento dos novos atributos gerados. Entretanto nessa análise não foi viável criar histogramas de frequência comparando todas as especialidades, pois há muitas especialidades.

Um caso similar ao ocorrido nos grupos de códigos de CID, no qual um atributo possui uma dimensionalidade extremamente maior que os demais atributos, é o atributo CLINICA MÉDICA que possui 132.557 ocorrências (43,36%) ante 173.166 ocorrências (56,64%) das demais especialidades. Portanto é conveniente descartá-la para construção de um histograma de frequência com todas as especialidades. Entretanto, como citado anteriormente, devido ao alto número de atributos, a construção do histograma tornou-se inviável e fez-se necessário analisar as frequências dos atributos por meio de uma tabela.

Na análise das frequências de especialidades percebeu-se atributos com baixa cobertura, o que os tornam inviáveis para MD uma vez que dificilmente fornecerão padrões de alta qualidade[6]. Também são analisadas as frequências desconsiderando repetições onde se percebeu uma diminuição da proporção entre o atributo CLINICA MÉDICA (29,04%) e as demais especialidades (70,96%).

SADT, Consultas e Internações

Segundo a Agência Nacional de Saúde[7], a Guia de Serviços Profissionais/ Serviço Auxiliar Diagnóstico e Terapia (SP/SADT) deve ser utilizada no atendimento a diversos tipos de eventos, tais como: remoção, pequena cirurgia, terapias, consulta com procedimentos, exames, atendimento domiciliar, SADT internado ou quimioterapia, radioterapia ou terapia renal substitutiva (TRS). A Guia de Consultas deve ser utilizada exclusivamente na execução de consultas eletivas sem procedimento, já a Guia de Solicitação de Internação é referente aos casos de internação. Essas informações estão presentes nos dados por meio dos atributos SADT, CONSULTAS e INTERNACÃO, que estão representadas na Figura 3.

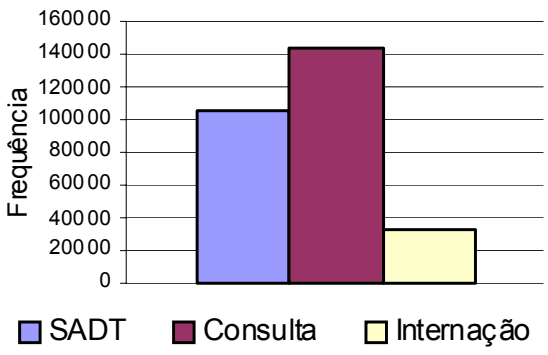

Figura 3 - Frequência SADT, Consulta e Internação

Fica claro no histograma da Figura 3 que o número de consultas é o maior entre os três procedimentos, e o número de internações é bem mais baixo que o número de consultas e SADT. Este histograma leva em consideração o número de vezes que um usuário utilizou cada serviço.

O histograma da Figura 4 demonstra a relação entre esses três atributos sem considerar repetições. Neste histograma é possível observar que a proporção de internações e SADT diminuem consideravelmente em relação a consultas o que pode nos levar a concluir que beneficiários que já tiveram uma internação estão propensos a sofrer novas internações.

Sexo, Idade, Cidade e Estado

A análise do atributo SEXO demonstrou que do total de beneficiários (159.883) a maior proporção (58,91%) é de beneficiários do sexo feminino (94.187 registros).

A idade dos beneficiários varia entre -1 e 105 anos com uma concentração maior em torno dos 50 anos. A idade -1 corresponde a um ruído nos dados e podem ser casos de idade não informada ou algum problema no processo de entrada de dados. Nestes casos estes registros foram desconsiderados.

A maioria dos beneficiários (85,74%) pertence a um mesmo estado. Os demais estados possuem quantidade irrisória e 95% dos 22794 beneficiários restantes não possuem informação de estados em seus cadastros. O atributo MUNICÍPIO possui um domínio com 356 cidades, com maior concentração em uma região metropolitana, composta por duas cidades, com 32301 beneficiários. Há também 14% do total de beneficiários que não possuem suas cidades informadas.

Grupo de eventos

Eventos são procedimentos realizados pelos beneficiários dentro ou fora dos hospitais, como por exemplo um medicamento que deve ser aplicado. O alto

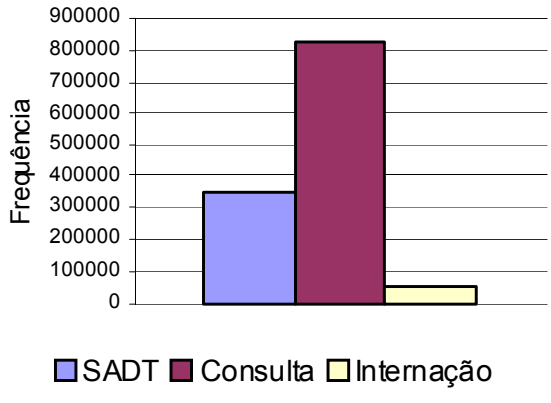

Figura 4 - Frequência SADT, Consulta e Internação sem repetições

número de eventos cadastrados na base tornou difícil a análise desses dados na forma como eram armazenados. Estudando a base de dados percebeu-se que esses eventos sofriam um agrupamento, ou seja, buscava-se agrupar eventos que são realizados por beneficiários com algum problema de saúde. Por exemplo, o agrupamento "procedimentos relacionados a câncer de mama" procura determinar quais eventos estão relacionados as pessoas que tenham câncer de mama.

Os histogramas das Figuras 5 e 6 demonstram as frequências desses agrupamentos de eventos.

Percebe-se pelo histograma da Figura 5 que alguns grupos não possuem ou quase não possuem dados o que inviabiliza a utilização dos mesmos para MD.

Quando se desconsidera as repetições, como no histograma da Figura 6, a proporção de dados cai mais permitindo a seleção de alguns grupos de eventos.

Seleção de atributos relevantes

Segundo Han e Kamber[8], a seleção de atributos em um conjunto de dados reduz o tamanho do conjunto removendo atributos irrelevantes ou redundantes. Embora seja possível para um especialista de domínio escolher alguns atributos úteis isso pode ser uma tarefa complicada e consumir muito tempo quando o comportamento dos dados não é conhecido. Deixar atributos relevantes fora do conjunto de atributos selecionados ou manter atributos irrelevantes pode causar confusão para o algoritmo de MD. Isto pode resultar na descoberta de padrões de baixa qualidade. Além disso, o alto número de atributos irrelevantes ou redundantes aumenta o custo computacional do processo de DCBD.

Um conceito importante que foi levado em consideração, para seleção dos atributos, foi o conceito de Prevalência de Classe[6] referente ao desbalanceamento de classes. Por exemplo, considere um conjunto com a seguinte distribuição de classes **dist** (C_1,C_2,C_3) = (99,00%, 0,25%, 0,75%), com prevalência da classe C_1. Um classificador simples que classifique sempre novos exemplos como pertencentes à classe majoritária C_1 teria uma precisão de 99,00%. Isto pode ser indesejável quando as classes minoritárias possuem alguma informação muito importante. Por exemplo, supondo C_1: paciente normal, C_2: paciente com doença A e C_3: paciente com doença B. Isso ocorre por exemplo no GRUPO-CID-VIIIC (Doenças crônicas das vias aéreas inferiores) que possui uma distribuição de classes **dist**(SIM, NÃO) = (1,30%, 98,70%), em que SIM: paciente com doença e NÃO: paciente normal, com prevalência da classe NÃO. No caso deste atributo, o número de registros da classe SIM é igual a 2071, valor pouco significativo no universo de 159883 registros deste conjunto de dados. Portanto, este atributo deve ser eliminado do conjunto de atributos selecionados. Este conceito foi aplicado na seleção de todos os atributos.

Com base nas análises das seções anteriores e o

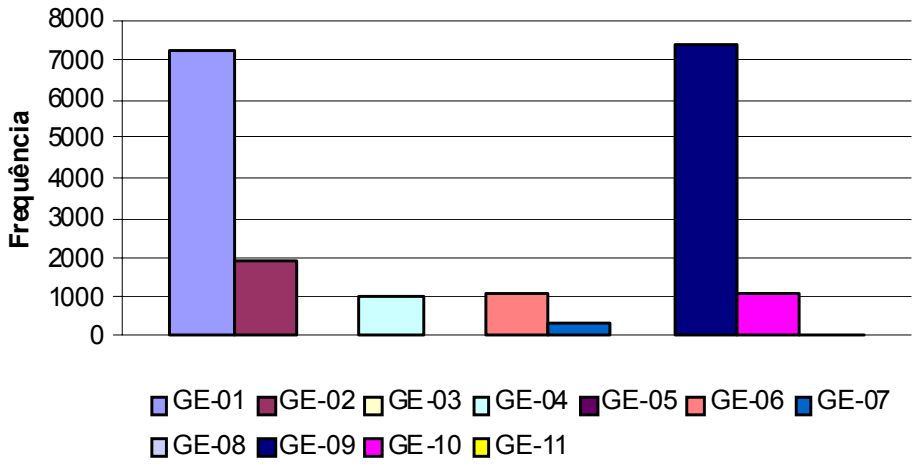

Figura 5 - Frequência Grupos de Eventos

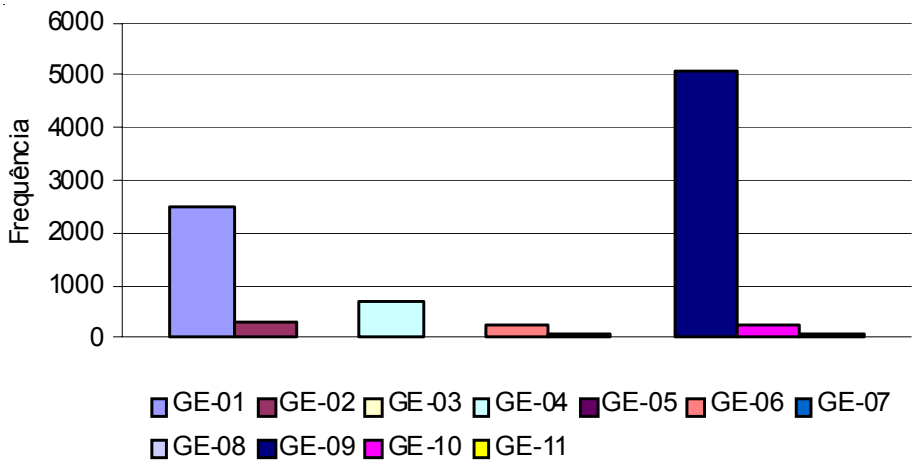

Figura 6 - Frequência Grupos de Eventos sem repetições

conceito de prevalência de classe fez-se a seleção dos atributos para MD. Primeiramente fez-se a seleção dos grupos de CIDs, com base nos histogramas da Figura 2 e 4, selecionou-se os atributos com maior frequência da classe SIM e descartou-se os atributos com frequência menor que 2500 registros da classe SIM, devido a diferença desproporcional relativa aos demais atributos. O mesmo procedimento foi efetuado para especialidade, mas foram eliminadas as especialidades com frequência menor que 700 ocorrências da classe SIM. Os atributos SADT, Consultas e Internações foram mantidas pela sua relevância e alto número de ocorrências, já os atributos Sexo, Idade, Cidade e Estado também foram mantidos por possuírem informações importantes em relação aos beneficiários. Em relação aos grupos de eventos selecionou-se os grupos 01 (Procedimentos relacionados a Diabetes Mellitus), 04 (Procedimentos Sentinelas relacionadas a Diabetes Meliltus) e 09 (Procedimentos relacionados ao Câncer de Mama) por possuírem alta frequência.

RESULTADOS

Com o objetivo de obter dados preparados para MD, aplicou-se a metodologia descrita neste trabalho para selecionar os principais atributos, além de eliminar os registros inúteis ou com problemas nos dados. No apêndice A encontra-se uma tabela contendo os atributos selecionados.

Com a aplicação desta metodologia obteve-se uma redução de 120 para 55 atributos e redução em 14% no número de registros considerados. Obtiveram-se também dados de melhor qualidade que serão disponibilizados para utilização por algoritmos de MD desenvolvidos pelo GPEA, além de outros pesquisadores que queiram utilizar esses dados. A forma como esses dados estão organizados agora facilita a aplicação de algoritmos de MD e ainda podem sofrer algum processo de refinamento dependendo do foco que se queira dar a utilização desses dados.

CONCLUSÃO

Demonstrou-se o pré-processamento de dados sobre beneficiários de um plano de saúde suplementar. Houve algumas dificuldades para compreensão do domínio da aplicação, as quais foram superadas por meio da análise da estrutura de dados e estudo da documentação da estrutura, obtendo-se informações como por exemplo o significado de tabelas e atributos além de conhecimentos referentes aos processos de saúde suplementar. Utilizou-se também análise estatística baseada no uso de tabelas e gráficos.

Com os estudos feitos sobre os dados foi possível obter uma visão mais clara do domínio da aplicação. Essa visão facilitou a seleção de atributos e o entendimento de alguns comportamentos como, por exemplo, casos em que pacientes que sofrem internação estão mais propensos a sofrerem novas internações. Isto foi constatado por meio dos gráficos de frequência.

A utilização de gráficos facilitou a escolha dos atributos e ajudou na eliminação de atributos de baixa frequência, que provocaria a descoberta de regras de baixa qualidade devido a baixa cobertura desses atributos, o que ocasionaria um grande esforço computacional desnecessário na aplicação dos algoritmos de MD pois aumentaria o espaço de busca.

Esta base de dados preparada facilitará a exploração de padrões ocultos por algoritmos de MD economizando tempo precioso dos pesquisadores que utilizarem esses dados uma vez que a etapa de pré-processamento, a qual consome mais tempo dentro do processo de DCBD, já está pronta. Além disso, a redução obtida no número de atributos e no número de registros implicou na redução significativa no espaço de busca que irá proporcionar uma expressiva redução no tempo computacional necessário para descoberta de conhecimento relevante.

A metodologia de pré-processamento desenvolvida neste trabalho poderá ser aplicada a outras bases de dados sobre planos de saúde suplementar.

REFERÊNCIAS

1. Freitas AA. Data mining and knowledge discovery with evolutionary algorithms. Canterbury: Stringer; c2002. 264 p.
2. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996; 17(3): 37-54.
3. Neves RC. Pré-processamento no processo de descoberta de conhecimento em banco de dados [Dissertação]. Porto Alegre: Instituto de Informática, Universidade Federal do Rio Grande do Sul; 2003.
4. Castro MS, Carvalho MS. Agrupamento da classificação internacional de doenças para análise de reinternações hospitalares. Cad Saúde Pública 2005;21(1): 317-23.
5. Brasil. Ministério da Saúde. DATASUS - Departamento de Informática do SUS. Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde - CID-10. [citado 2011 fev 11]. Disponível em: http://www.datasus.gov.br/cid10/v2008/webhelp/z00_z13.htm#Z00.
6. Monard MC, Baranauskas JA. Conceitos sobre aprendizado de máquina. In: Rezende SO. Sistemas inteligentes fundamentos e aplicações. Barueri: Manole; 2005. p.89-114.
7. Brasil. Ministério da Saúde. Agência Nacional de Saúde Suplementar. Guia de serviços profissionais/Serviço Auxiliar Diagnóstico e Terapia. [citado 2000 jan. 1]. Disponível em: http://www.ans.gov.br/portal/site/_hotsite_tiss/padrao_tiss.htm#spsadt1.
8. Han J, Kamber M. Data mining concepts and techniques. In: Kaufmann M. Data preprocessing. 2nd ed. London: Oxford; 2006. p.47-103.

Apêndice A – Tabela final com os atributos selecionados

Atributo	Domínio	Descrição
Sexo	F, M	Sexo do beneficiário
Idade	[0..105]	Idade do beneficiário
Estado	SC, RS, PR, SP, RN, MG, RJ, DF, AL, MT, GO	Estado de residência do beneficiário
Município	[ABDON BATISTA.ZORTEA]	Município de residência do beneficiário
SADT	[0..285]	Quantidade de Serviço Auxiliar de Diagnóstico e Terapia Executado
Consulta	[0..101]	Quantidade de consultas médicas realizadas
Internação	[0..63]	Quantidade de internações hospitalares realizadas
Grupo-CID-I	[0..69]	Doenças infecciosas e parasitárias
Grupo-CID-II	[0..95]	Neoplasias
Grupo-CID-IIIA	[0..41]	Doenças endócrinas
Grupo-CID-IIIB	[0..49]	Doenças nutricionais e metabólicas
Grupo-CID-V	[0..272]	Doenças mentais
Grupo-CID-VIA	[0..175]	Sistema nervoso
Grupo-CID-VIB	[0..80]	Olhos e anexos, ouvido e apófise mastóide
Grupo-CID-VIIA	[0..144]	Circulatório, menos veias e linfáticos
Grupo-CID-VIIIA	[0..74]	Infecções respiratórias agudas
Grupo-CID-VIIIB	[0..64]	Outras doenças respiratórias
Grupo-CID-IXA	[0..138]	Doenças do esôfago, estômago e duodeno
Grupo-CID-XII	[0..48]	Pele e tecido subcutâneo
Grupo-CID-XIII	[0..115]	Osteomuscular e tecido conjuntivo
Grupo-CID-XA	[0..56]	Doenças urinárias
Grupo-CID-XB	[0..85]	Doenças genitais masculinas
Grupo-CID-XC	[0..117]	Doenças genitais femininas
Grupo-CID-XVI	[0..93]	Sintomas, sinais e afecções mal definidas
Grupo-CID-XVII	[0..82]	Lesões, envenenamentos e causas externas
Grupo-CID-XXI	[0..66]	Contato com os serviços de saúde
Especialidade-04	[0..16]	Análises clínicas
Especialidade-05	[0..8]	Anatomia patológica
Especialidade-06	[0..6]	Anestesioliga
Especialidade-07	[0..7]	Angiologia
Especialidade-09	[0..16]	Cardiologia
Especialidade-13	[0..8]	Cirurgia geral
Especialidade-18	[0..56]	Clínica médica
Especialidade-24	[0..7]	Dermatologia
Especialidade-27	[0..4]	Endocrinologia
Especialidade-31	[0..5]	Gastroenterologia
Especialidade-33	[0..12]	Ginecologia
Especialidade-39	[0..23]	Internações
Especialidade-47	[0..12]	Medicina preventiva
Especialidade-50	[0..5]	Neurologia
Especialidade-53	[0..9]	Oftalmologia
Especialidade-54	[0..11]	Ortopedia
Especialidade-55	[0..9]	Otorrinolaringologia
Especialidade-56	[0..6]	Patologia
Especialidade-57	[0..11]	Patologia clínica
Especialidade-58	[0..13]	Pediatria
Especialidade-59	[0..6]	Pneumologia
Especialidade-62	[0..6]	Psiquiatria
Especialidade-63	[0..5]	Radiodiagnostico
Especialidade-64	[0..9]	Radiologia
Especialidade-66	[0..4]	Reumatologia
Especialidade-69	[0..9]	Urologia
Grupo-Evento-01	[0..178]	Procedimentos – Diabetes Mellitus
Grupo-Evento-04	[0..19]	Procedimentos sentinelas – Diabetes Mellitus
Grupo-Evento-09	[0..45]	Procedimentos – câncer de mama