

UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE ELETRICIDADE
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

JOCIELMA JERUSA LEAL ROCHA

**APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA
PARA CLASSIFICAÇÃO DE DEPÓSITOS MINERAIS BASEADA
EM MODELO TEOR-TONELAGEM**

São Luís

2010

JOCIELMA JERUSA LEAL ROCHA

**APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA
PARA CLASSIFICAÇÃO DE DEPÓSITOS MINERAIS BASEADA
EM MODELO TEOR-TONELAGEM**

Dissertação de Mestrado submetida à
Coordenação do Programa de Pós-Graduação
em Engenharia de Eletricidade da Universidade
Federal do Maranhão como parte dos requisitos
para obtenção do título de Mestre em Ciência da
Computação.

Orientador: Prof. Dr. Anselmo Cardoso de Paiva

Co-orientador: Prof. Dr. Aristófanés Corrêa Silva

São Luís

2010

Rocha, Jocielma Jerusa Leal

Aplicação de Técnicas de Aprendizado de Máquina para Classificação de Depósitos Minerais Baseada em Modelo Teor-Tonelagem / Jocielma Jerusa Leal Rocha. - São Luís, 2010.

108f. : il.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Maranhão, 2010.

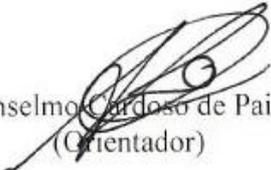
1. Aprendizado de Máquina. 2. Classificação de Depósitos Minerais. 3. Modelo Teor-Tonelagem. 4. Combinação de Classificadores. I. Título

CDU 004.85

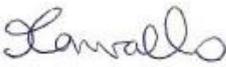
**APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA
PARA CLASSIFICAÇÃO DE DEPÓSITOS MINERAIS
BASEADA EM MODELO TEOR-TONELAGEM**

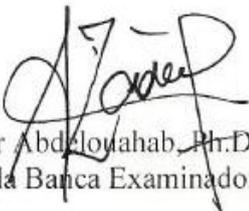
Jocielma Jerusa Leal Rocha

Dissertação aprovada em 01 de julho de 2010.


Prof. Anselmo Cardoso de Paiva, Dr.
(Orientador)


Prof. Aristófares Corrêa Silva, Dr.
(Co-orientador)


Prof. Paulo Cezar Pinto Carvalho, Dr.
(Membro da Banca Examinadora)


Prof. Zair Abdelouahab, Ph.D.
(Membro da Banca Examinadora)

*“Nosso maior desejo na vida é encontrar
alguém que nos faça fazer o melhor que
pudermos.”*

Ralph Waldo Emerson

À minha família.

AGRADECIMENTOS

Ao meu Senhor e meu Deus, pelo seu amor incondicional.

À minha família, pelo apoio, compreensão e toda a felicidade que me proporcionam com sua existência.

Aos meus orientadores, pelo incentivo, paciência e ensinamentos valorosos.

Aos pesquisadores que colaboraram com o material inicial de depósitos minerais e com a base de dados utilizada neste trabalho, em especial ao Sr. Donald Singer, pela disponibilidade colocada para esclarecimento de dúvidas e discussão.

Aos professores e colegas de laboratório, pela ajuda e ânimo pela pesquisa.

Aos colegas de trabalho, pelo respeito e incentivo de suas palavras.

Aos meus amigos, pela compreensão nos momentos em que me ausentei de sua companhia e alegria nos momentos em que estive presente.

Aos colegas que me ajudaram com as figuras quando eu já não tinha mais paciência.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

RESUMO

A classificação de depósitos minerais em tipos tradicionalmente é feita por especialistas no assunto. A possibilidade de que técnicas computacionais auxiliem o processo de classificação e o torne menos subjetivo incentiva a pesquisa e aplicação de diferentes métodos de agrupamento e classificação sobre esse domínio de análise. A evolução das pesquisas nesse domínio tem direcionado os estudos para a utilização de informações disponíveis em grandes bases de dados publicadas e a aplicação de técnicas de aprendizado de máquina supervisionado. Este trabalho utiliza informações de depósitos minerais disponibilizadas em modelos teor-tonelagem publicados na literatura para proceder a investigação da adequabilidade de três dessas técnicas: Árvore de Decisão, Rede Percéptron Multicamadas e Rede Neural Probabilística. Ao todo, são 1.861 depósitos distribuídos em 18 tipos identificados pelo modelo teor-tonelagem. Inicialmente verificou-se o resultado apresentado por cada uma das três técnicas para a classificação dos depósitos em 18 tipos. A análise desses resultados sugeriu a possibilidade de agrupar esses tipos e dividir a classificação em dois níveis: o primeiro nível para classificar os depósitos considerando o agrupamento de tipos e o segundo nível para classificar os depósitos que resultaram em um grupo em um dos tipos específicos daquele grupo. Uma série de experimentos foi realizada no sentido de construir um modelo de classificação em dois níveis a partir da combinação das técnicas utilizadas, o que resultou em uma taxa de acerto média de 85% dos casos e as principais ocorrências de erros foram identificadas dentro de grupos em tipos de depósitos menos representativos na base de dados. Isso representa uma maneira promissora de conseguir melhoria no processo de classificação de depósitos minerais que não implica no aumento da quantidade de depósitos utilizada ou na quantidade de características dos depósitos.

Palavras-chave: Classificação de Depósitos Minerais, Modelo Teor-Tonelagem, Aprendizado de Máquina Supervisionado, Combinação de Classificadores.

ABSTRACT

Classification of mineral deposits into types is traditionally done by experts. Since there are reasons to believe that computational techniques can aid this classification process and make it less subjective, the research and investigation of different methods of clustering and classification to this domain may be appropriate. The way followed by researches in this domain has directed for the use of information available in large public databases and the application of supervised machine learning techniques. This work uses information from mineral deposits available in grade-tonnage models published in the literature to conduct research about the suitability of these three techniques: Decision Tree, Multilayer Perceptron Network and Probabilistic Neural Network. Altogether, 1,861 mineral deposits of 18 types are used. The types refer to grade-tonnage models. Initially, each of these three techniques are used to classify mineral deposits into 18 types. Analysis of these results suggested that some deposits types could be treated as a group and also that the classification could be divided into two levels: the first level to classify deposits considering groups of deposits and the second level to classify deposits previously identified on a group into some of specific type belonging to that group. A series of experiments was carried out in order to build a two levels model from the combination of the techniques used, which resulted in an average accuracy rate of 85% of cases. Patterns of errors occurrence were identified within groups in types of deposits less representative in the database. This represents a promising way to achieve improvement in the process of mineral deposits classification that does not mean increasing in the amount of deposits used or in the amount of characteristics of the deposits.

Keywords: Mineral Deposits Classification, Grade-Tonnage Model, Supervised Machine Learning, Combination of Classifiers.

LISTA DE TABELAS

Tabela 3.1 – Quantidade de depósito por tipo na base de dados.	35
Tabela 3.2 - Distribuição dos depósitos da base de dados no conjunto de treinamento e no conjunto de teste.	43
Tabela 3.3 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais em 18 tipos.....	44
Tabela 3.4 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas para classificação em 18 tipos de depósitos.	46
Tabela 3.5 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais em 18 tipos.	47
Tabela 3.6 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística para classificação em 18 tipos de depósitos.	48
Tabela 3.7 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais em 18 tipos.....	49
Tabela 3.8 – Grupo de tipos de depósitos considerados.....	53
Tabela 3.9 – Distribuição dos depósitos agrupados no conjunto de treinamento e no conjunto de teste.	54
Tabela 3.10 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais considerando a existência de grupos de depósitos.....	55
Tabela 3.11 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando o agrupamento de tipos de depósitos.	56

Tabela 3.12 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais considerando a existência de grupos de depósitos.	57
Tabela 3.13 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando o agrupamento tipos de depósitos.	58
Tabela 3.14 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais considerando a existência de grupos de depósitos.	59
Tabela 3.15 – Distribuição dos depósitos do Grupo 1 no conjunto de treinamento e no conjunto de teste.	63
Tabela 3.16 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 1.	63
Tabela 3.17 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 1.	64
Tabela 3.18 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 1.	64
Tabela 3.19 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 1.	64
Tabela 3.20 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 1.	65
Tabela 3.21 – Acerto médio obtido em cada aplicação para depósitos do Grupo 1.	66
Tabela 3.22 – Distribuição dos depósitos do Grupo 6 no conjunto de treinamento e no conjunto de teste.	67

Tabela 3.23 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 6.	67
Tabela 3.24 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 6.	68
Tabela 3.25 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 6.	68
Tabela 3.26 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 6.	68
Tabela 3.27 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 6.	69
Tabela 3.28 – Acerto médio obtido em cada aplicação para depósitos do Grupo 6.	71
Tabela 3.29 – Distribuição dos depósitos do Grupo 9 no conjunto de treinamento e no conjunto de teste.	71
Tabela 3.30 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 9.	72
Tabela 3.31 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 9.	72
Tabela 3.32 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 9.	73
Tabela 3.33 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 9.	73

Tabela 3.34 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 9.	73
Tabela 3.35 – Acerto médio obtido em cada aplicação para depósitos do Grupo 9.....	75
Tabela 3.36 – Distribuição dos depósitos do Grupo 11 no conjunto de treinamento e no conjunto de teste.	75
Tabela 3.37 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 11.	76
Tabela 3.38 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 11.	76
Tabela 3.39 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 11.....	77
Tabela 3.40 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 11.	77
Tabela 3.41 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 11.	77
Tabela 3.42 – Acerto médio obtido em cada aplicação para depósitos do Grupo 11.....	79
Tabela 3.43 – Distribuição dos depósitos do Grupo 14 no conjunto de treinamento e no conjunto de teste.	79
Tabela 3.44 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 14.	80
Tabela 3.45 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 14.	80

Tabela 3.46 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 14.....	80
Tabela 3.47 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 14.	81
Tabela 3.48 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 14.	81
Tabela 3.49 – Acerto médio obtido em cada aplicação para depósitos do Grupo 14.....	82
Tabela 3.50 – Distribuição dos depósitos do Grupo 17 no conjunto de treinamento e no conjunto de teste.	83
Tabela 3.51 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 17.	83
Tabela 3.52 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 17.	84
Tabela 3.53 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 17.	84
Tabela 3.54 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 17.	84
Tabela 3.55 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 17.	85
Tabela 3.56 – Acerto médio obtido em cada aplicação para depósitos do Grupo 17.....	86

Tabela 3.57 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de classificação em dois níveis proposto no Experimento 9 para classificação de depósitos minerais em 18 tipos.....	90
Tabela 3.58 – Taxa de acerto para os modelos de classificação que diferenciam os depósitos em 18 tipos para comparação dos resultados dos experimentos.	91
Tabela 3.59 – Melhor aplicação para os tipos de depósitos que foram agrupados para comparação dos resultados dos experimentos.....	93
Tabela 3.60 – Melhor aplicação para os tipos de depósitos que não foram agrupados para comparação dos resultados dos experimentos.	94
Tabela 3.61 – Taxa de acerto para os modelos de classificação que diferenciam os depósitos em 18 tipos para comparação com resultados publicados.....	95
Tabela 3.62 – Melhor aplicação para os tipos de depósitos que foram agrupados para comparação com resultados publicados.	97
Tabela 3.63 – Melhor aplicação para os tipos de depósitos que não foram agrupados para comparação com resultados publicados.	97

LISTA DE FIGURAS

Figura 2.1 – A organização de informações em modelos de depósitos minerais. Fonte: Adaptada de (SINGER, 1996a).	10
Figura 2.2 – Representação de uma situação hipotética para construção de modelo de densidade de depósitos: (a) 12 regiões bem exploradas com suas respectivas densidades e (b) histograma das densidades de depósito derivadas. Fonte: Adaptada de (SINGER <i>et al</i> , 2001).	12
Figura 2.3 – Esquema do sistema de avaliação em três partes. Fonte: Adaptada de (SINGER, 1996a).	14
Figura 2.4 – Representação de uma árvore de decisão binária: (a) estrutura da árvore e (b) sua respectiva representação no espaço de atributos. Fonte: Adaptada de (GAMA, 2002).	20
Figura 2.5 – Rede Percéptron Multicamadas. Fonte: (COSTA, 2006).	27
Figura 2.6 – Rede Neural Probabilística. Fonte: Adaptada de (GEORGIU <i>et al</i> , 2004).	32
Figura 3.1 – Metodologia para comparação dos classificadores.	37
Figura 3.2 – Modelo de classificação formado pela combinação de classificadores organizados em 2 níveis de classificação.	88

LISTA DE GRÁFICOS

Gráfico 3.1 – Porcentagem de acertos dos classificadores considerando os 18 tipos de depósito.....	51
Gráfico 3.2 – Porcentagem de acertos dos classificadores considerando o agrupamento de depósitos.	61
Gráfico 3.3 - Porcentagem de acertos dos classificadores para o Grupo 1.	65
Gráfico 3.4 - Porcentagem de acertos dos classificadores para o Grupo 6.	70
Gráfico 3.5 - Porcentagem de acertos dos classificadores para o Grupo 9.	74
Gráfico 3.6 - Porcentagem de acertos dos classificadores para o Grupo 11.	78
Gráfico 3.7 - Porcentagem de acertos dos classificadores para o Grupo 14.	82
Gráfico 3.8 - Porcentagem de acertos dos classificadores para o Grupo 17.	86
Gráfico 3.9 – Comparativo entre classificadores gerados com apenas uma técnica de aprendizado de máquina e sem considerar agrupamento de tipos e o modelo que considera dois níveis de classificação em razão do agrupamento de tipos de depósitos.....	92
Gráfico 3.10 – Comparativo do resultado obtido com o modelo de classificação em dois níveis construído neste trabalho <i>versus</i> resultados publicados na literatura.	96

LISTA DE LISTAGENS

Listagem 3.1 – Algoritmo para realização dos experimentos.	37
Listagem 3.2 – Algoritmo para determinação dos valores dos parâmetros adequados para construção de modelos de Rede Percéptron Multicamadas de um experimento.	40
Listagem 3.3 – Algoritmo para determinação do valor do parâmetro adequado para construção de modelos de Rede Neural Probabilística de um experimento.....	42
Listagem 3.4 – Algoritmo para realização do Experimento 9.....	89

SUMÁRIO

1	Introdução	1
1.1	Trabalhos correlatos	4
1.2	Objetivos do trabalho	6
1.3	Organização da Dissertação.....	7
2	Fundamentação teórica	8
2.1	Conceitos Básicos sobre a Prospecção Mineral	8
2.2	Modelos de Depósitos Minerais	9
2.3	Avaliação Quantitativa de Recursos Minerais	13
2.4	Classificação de Depósitos Minerais.....	16
2.5	Árvores de Decisão.....	19
2.5.1	Crescimento da Árvore	21
2.5.2	Poda da Árvore	23
2.6	Rede Percéptron Multicamadas	25
2.7	Rede Neural Probabilística	28
2.7.1	Base Teórica	29
2.7.2	Arquitetura.....	31
2.7.3	Fluxo da Rede Neural Probabilística	32
3	Experimentos e Resultados	34
3.1	Dados	34
3.2	Metodologia dos Experimentos	36
3.2.1	Modelos de Árvore de Decisão.....	39
3.2.2	Modelos de Rede Percéptron Multicamadas	39
3.2.3	Modelos de Rede Neural Probabilística	41

3.3	Experimento 1	42
3.4	Experimento 2	53
3.5	Experimento 3	62
3.6	Experimento 4	66
3.7	Experimento 5	71
3.8	Experimento 6	75
3.9	Experimento 7	79
3.10	Experimento 8	83
3.11	Experimento 9	87
3.12	Comparação dos Resultados dos Experimentos	91
3.13	Comparação com Resultados Publicados	95
4	Conclusão	99
5	Referências bibliográficas	102

1 Introdução

O crescimento da atual sociedade industrial e de consumo carece sobremaneira de bens econômicos minerais, de modo que o desenvolvimento de melhores técnicas para a descoberta de depósitos minerais é almejado por diversas instituições de pesquisa no mundo. Ao conjunto de conhecimentos que permitem o planejamento e a execução de trabalhos para descoberta de depósitos minerais denomina-se prospecção mineral (PEREIRA, 2003).

A execução das atividades relacionadas à descoberta e caracterização de jazidas minerais é bastante onerosa e também é muito grande a probabilidade de insucesso, pois nem toda área com indícios de mineralização é interessante para ser explorada economicamente. Assim, em vários momentos as atividades de prospecção são diretamente dependentes da análise da viabilidade custo-risco para seu desenvolvimento. É papel do prospector verificar as melhores técnicas e metodologias, na busca por otimizar a campanha de prospecção em termos de custos, tempo e recursos. Segundo Maranhão (MARANHÃO, 1989), dois princípios norteiam suas decisões: o princípio da analogia e o princípio das aproximações sucessivas.

O princípio da analogia incentiva a elaboração de modelos de depósitos minerais pela justificativa de que depósitos formados em condições geológicas similares apresentam muitas características em comum, mesmo que ocorram em diferentes países ou continentes. O princípio das aproximações sucessivas refere-se à necessidade de aplicação de modelos em cada etapa do projeto, até que se disponha de informações suficientes para representar apropriadamente o caso específico em análise. Os princípios descritos ressaltam, então, a importância da elaboração e do aperfeiçoamento de modelos de depósitos minerais para a condução sistemática dos trabalhos de busca por recursos minerais.

Segundo (PEREIRA, 2003) para que uma campanha de prospecção mineral de uma região seja plenamente atendida é indispensável a execução de três etapas: exploração geológica, prospecção de superfície e avaliação dos depósitos. A primeira etapa, a exploração geográfica, visa a identificação de áreas prioritárias para um estudo mais detalhado. As áreas identificadas são alvo de estudos detalhados na etapa de prospecção de superfície. A etapa de avaliação de depósitos é responsável pela

quantificação dos depósitos quanto aos recursos e teores minerais. Essas etapas são sucessivas e dependentes, de modo que ao fim de cada etapa deve-se verificar a viabilidade de continuação ou encerramento dos trabalhos na região afim de que seja possível reduzir os riscos de investimentos na pesquisa de recursos minerais em regiões que se revelem impróprias para exploração.

O acesso a informações sobre recursos não descobertos torna-se importante nesse contexto, pois fundamenta a tomada de decisões sobre a viabilidade dos trabalhos. Tais informações podem ser obtidas através de avaliações quantitativas de recursos minerais não descobertos. No entanto, por sua própria natureza, as avaliações refletem grandes incertezas. A partir da identificação das fontes de tais incertezas, estas podem ser reduzidas, bem como os riscos ocasionados por elas.

Uma maneira muito utilizada para realizar avaliação quantitativa de recursos não descobertos é o sistema de avaliação em três partes descrito em (SINGER and COX, 1988; SINGER, 1993). As três partes desse sistema são: (1) a delimitação de áreas em que a geologia permite a ocorrência de determinados tipos de depósitos; (2) a geração de estimativas sobre o número de depósitos não descobertos nas áreas para cada tipo de depósito; (3) a geração de estimativas sobre as quantidades de minério e sobre algumas características dos minerais dessas áreas. Essa metodologia destaca o uso consistente de modelos de depósitos, tais como modelos descritivos, modelos teor-tonelagem e modelos de densidade de depósitos, para a geração de estatísticas sobre depósitos não descobertos com um mínimo de influência que possa tornar os resultados tendenciosos.

Modelos descritivos identificam o ambiente geológico e configurações tectônicas das regiões onde pode ocorrer um tipo de depósito mineral e as características essenciais dos depósitos para esse tipo e, dessa forma, podem ser usados no processo de delimitação de áreas alvo. O modelo teor-tonelagem agrupa informações sobre a distribuição da frequência dos teores e tonelagens de depósitos de um determinado tipo e são utilizados para estimar teores e tonelagem de depósitos desconhecidos. Modelos de densidade de depósitos procuram representar a densidade, ou seja, a distribuição do número de depósitos minerais de um determinado tipo por uma medida de área e podem ser utilizados para estimar o número de depósitos não descobertos.

A consistência das avaliações pelo sistema de avaliação em três partes está intimamente relacionada à existência de modelos teor-tonelagem para os tipos de depósitos identificados, pois as áreas são delineadas de acordo com modelos descritivos com os quais os modelos teor-tonelagem são consistentes e as estimativas do número de depósitos devem ser consistentes com o modelo teor-tonelagem. Em (SINGER, 1993) afirma-se que as influências introduzidas nesse sistema provêm de duas fontes: da utilização de modelo teor-tonelagem incorreto ou da falta de consistência na estimativa do número de depósitos com o modelo teor-tonelagem. Aplicações do sistema de avaliação em três partes devem procurar utilizar modelos gerais como os descritos em (BLISS, 1992; COX and SINGER, 1986), a menos que as características dos depósitos encontrados na região em análise sejam muito diferentes dos modelos gerais, caso em que seria necessária a elaboração de modelos específicos para a região (SINGER, 1996b).

É possível falar na importância da classificação de depósitos minerais em tipos para avaliação quantitativa de recursos minerais. De fato, pelo exposto, a identificação correta do tipo de depósito implica na seleção do modelo teor-tonelagem correto. Por sua vez, a escolha correta do modelo teor-tonelagem tem conseqüências tanto na confiabilidade das estimativas dos teores e tonelagens dos depósitos da região quanto na garantia da consistência do sistema de avaliação em três partes, o qual aplica o modelo de densidade de depósitos para estimar o número de depósitos não descobertos. Segundo (SINGER, 2006) a classificação de depósitos e ocorrências em tipos auxilia o processo de delimitação e estimação do número de depósitos não descobertos e, em alguns casos, os tipos de depósitos e ocorrências conhecidos identificam ambientes geológicos não indicados em mapas geológicos.

Normalmente, a classificação de depósitos em tipos é feita por especialistas no assunto. É possível que técnicas computacionais e de aprendizado de máquina acrescentem valor a essa classificação e a torne menos subjetiva.

Problemas relacionados à classificação de dados podem ser resolvidos pela aplicação de técnicas de aprendizado de máquina supervisionado. Nesse caso, fala-se no treinamento de algoritmos através de um conjunto de dados cujas características e classes são conhecidas na intenção de gerar um classificador que permita inferir, com

aceitável nível de confiança, uma classe para dados que tem somente as suas características conhecidas.

O problema de identificar o modelo teor-tonelagem correto pode ser pensado como o problema de classificar corretamente os depósitos da região em estudo em tipos, em que cada tipo possui um modelo teor-tonelagem associado. Desse modo, deve ser selecionado um conjunto de características dos depósitos e um grupo de classes às quais estes podem estar associados. As classes dos depósitos seriam os tipos de depósitos que possuem um modelo teor-tonelagem associado. As características poderiam estar relacionadas à descrição de depósitos em modelos teor-tonelagem.

Algumas técnicas de aprendizado de máquina podem se mostrar mais adequadas quando aplicadas a um determinado conjunto de dados. Assim, é interessante que diferentes métodos sejam aplicados para verificar a possibilidade de classificação automatizada de depósitos minerais.

1.1 Trabalhos correlatos

Um dos primeiros esforços para propiciar a avaliação quantitativa de depósitos minerais é encontrado em (BARTON, 1986) com a publicação de estimativas relacionadas à frequência de ocorrências minerais por tipo de depósitos. Tais frequências foram obtidas a partir da verificação da ocorrência de cerca de 150 minerais em aproximadamente 80 tipos de depósitos. Tal publicação constitui um passo fundamental para a construção de modelos quantitativos, os quais auxiliam a análise da viabilidade de investimentos em pesquisas e do financiamento da exploração mineral de uma região. Em (MCCAMMON, 1992), essas frequências foram utilizadas em conjunto com estimativas subjetivas da frequência de tipos de rochas associadas, idades geológicas, alterações, análises geofísicas e geoquímicas para classificação de depósitos minerais conhecidos por meio de um sistema especialista – o Prospector II. Esses testes resultaram num acerto de 83% na classificação de 124 depósitos do Alaska.

Os trabalhos encontrados que cronologicamente sucedem a aplicação de sistema especialista relatam a aplicabilidade de Redes Neurais a esse domínio de classificação (SINGER and KOUDA, 1996; SINGER and KOUDA, 1999b). Uma Rede

Neural Probabilística foi aplicada a dados sobre a presença ou ausência de 58 minerais para classificação de depósitos conhecidos em 8 tipos gerais de depósitos (SINGER and KOUDA, 1997a). Tal aplicação resultou em 88% de depósitos corretamente classificados. A adição da indicação da presença ou ausência de dois tipos de rochas representou um aumento para 98% da taxa de acerto para aquela pesquisa, o que sugeriu a continuidade da pesquisa com novos testes com Redes Neurais Probabilísticas considerando outros tipos de depósitos e a adição de novas características.

Novamente, em (SINGER and KOUDA, 2003), uma Rede Neural Probabilística foi utilizada, mas dessa vez foram utilizadas informações a respeito de teores e tonelagens de depósitos minerais. A mudança no tipo de informações foi justificada pela tentativa de utilizar informações acessíveis nos estágios iniciais do desenvolvimento de uma mina como são os teores de minérios e a tonelagem do depósito. Tal aplicação de Redes Neurais Probabilísticas a teores e tonelagens como atributos preditores do tipo de depósito obteve uma taxa de acerto de 75% quando apenas as médias dos teores e as tonelagens foram usadas para caracterização dos depósitos e, ainda, um acerto de 93% em outro teste com o acréscimo de informações sobre a presença ou ausência de 6 tipos de rochas e sobre mineralogia. Naqueles testes foram considerados 13 tipos de depósitos como resultados da classificação.

A partir da constatação de que teores e tonelagens poderiam ser usados para classificação dos depósitos e, ainda, de que informações sobre as rochas presentes no depósito melhoram os resultados obtidos, um novo estudo foi desenvolvido em (SINGER, 2006). Nele foram usadas como informações dos depósitos: a tonelagem, a média do teor dos minerais cobre (Cu), molibdênio (Mo), prata (Ag), ouro (Au), zinco (Zn) e chumbo (Pb), e presença de 6 tipos de rocha (*carbonates, felsic to intermediate volcanic, granitoids, mafic volcanic, marine felsic to intermediate volcanic, marine mafic volcanic*). Nesse caso, foram utilizados aproximadamente 1610 depósitos previamente classificados por especialistas no assunto entre 18 tipos de depósitos. A aplicação de Rede Neural Probabilística sobre esses dados obteve uma taxa de acerto de aproximadamente 82%. Uma verificação interessante desse artigo foi a análise da distribuição dos erros de classificação apresentada por alguns tipos de depósitos, como se formassem subgrupos, tal que, quando um depósito de um desses tipos era mal classificado, o resultado da classificação indicava um dos tipos do subgrupo. Nesse

caso, chegou-se a afirmar que a taxa de acerto do estudo era de 87% se depósitos do subgrupo *Porphyry copper* fossem agrupados.

Os trabalhos citados oferecem uma breve história da tentativa de utilização de métodos automatizados para a classificação de depósitos minerais. O caminho seguido incluiu experimentos realizados sobre diferentes conjuntos de informações dos depósitos e de algumas técnicas de classificação. A evolução das pesquisas tem direcionado os estudos para a utilização de informações disponíveis em grandes bases de dados de depósitos publicadas e para a aplicação de modelos de classificações baseados em técnicas de aprendizado de máquina (BEHNIA, 2007; SINGER and BERGER, 2007; LEITE e SOUZA FILHO, 2009; ZUO *et al*, 2009; SINGER, 2010). Este trabalho representa a tentativa de utilizar tais informações para classificar depósitos minerais através da aplicação de diferentes técnicas de aprendizado de máquina.

1.2 Objetivos do trabalho

Este trabalho visa investigar a adequabilidade de diferentes técnicas de aprendizado de máquina supervisionado para classificação de depósitos minerais baseada em modelo teor-tonelagem.

Os objetivos específicos deste trabalho são:

- Investigação da adequabilidade de um algoritmo baseado no método de aprendizado lógico, a Árvore de Decisão;
- Investigação da adequabilidade de um algoritmo baseado em técnicas de Redes Neurais Artificiais que possibilita a manipulação de problemas não linearmente separáveis, a Rede Percéptron Multicamadas;
- Investigação da adequabilidade de um algoritmo baseado em técnicas bayesianas, a Rede Neural Probabilística;
- Investigação da possibilidade de realizar uma classificação combinada através da construção de um modelo de classificação em dois níveis que considera os melhores resultados apresentados pelas técnicas anteriores para construir os classificadores que servem como base para o modelo.

1.3 Organização da Dissertação

Neste primeiro capítulo são apresentados os objetivos do trabalho, assim como as motivações para a pesquisa de técnicas de aprendizado de máquina adequadas à classificação de depósitos minerais em tipos. Além disso, são apresentados trabalhos anteriores sobre a classificação de depósitos minerais.

O Capítulo 2 apresenta conceitos que fundamentam o uso de modelo teor-tonelagem como base para a classificação, bem como as técnicas de aprendizado de máquina utilizadas.

A metodologia proposta para a investigação da aplicabilidade das técnicas é apresentada no Capítulo 3. Ele descreve os experimentos realizados, apresenta seus resultados e os discute através de comparações entre os resultados obtidos nos experimentos realizados neste trabalho e resultados publicados na literatura.

O Capítulo 4 apresenta as conclusões do trabalho e sugestões para continuidade da pesquisa.

2 Fundamentação teórica

A classificação de depósitos minerais em tipos torna-se especialmente interessante dentro do sistema de avaliação em três partes. A consistência requerida para avaliação de recursos minerais seguindo essa sistemática evidencia a importância da escolha correta de modelos a serem utilizados em cada etapa do processo de prospecção mineral. A identificação de tipos de depósitos possíveis em uma região direciona a escolha dos modelos. Para utilizar o sistema de avaliação em três partes é de salutar importância que esses tipos possam ser associados a modelos teor-tonelagem apropriados. Assim, a classificação dos depósitos e ocorrências minerais de uma região visando identificar o modelo teor-tonelagem correto torna-se importante para a prospecção mineral na medida em que a geração de estimativas sobre recursos não descobertos por essa sistemática torna-se mais confiável.

Este capítulo visa oferecer a fundamentação teórica necessária para compreensão da importância da utilização do modelo teor-tonelagem na classificação dos depósitos dentro do sistema de avaliação em três partes. Além disso, apresenta as técnicas utilizadas nesse trabalho para proceder à classificação. Assim, inicialmente são apresentados conceitos relacionados ao domínio da prospecção mineral e em seguida, conceitos relacionados às técnicas de aprendizado de máquina utilizadas.

2.1 Conceitos Básicos sobre a Prospecção Mineral

A existência de depósitos minerais está relacionada à evolução da crosta terrestre de tal modo que a distribuição de tipos de depósitos está condicionada a determinados fatores ou características que atualmente são descritos em controles de mineralização (MARANHÃO, 1989). Referem-se ao ambiente tectônico, tipo litológico, período geológico de formação das rochas e demais características que condicionam a existência de depósitos. Depósitos específicos podem apresentar características comuns, por exemplo, em relação aos minerais que possuem e ao ambiente onde ocorrem e, dessa forma, podem ser agrupados em tipos (COX and SINGER, 1986). Assim, a busca por recursos minerais orienta-se em um primeiro momento por controles de mineralização que podem ser associados aos tipos de depósitos.

A pesquisa ou prospecção mineral foi definida em (BATISTA, 2006) como o conjunto de operações e atividades empreendidas, sem garantia de sucesso, para evidenciar a existência de uma jazida, visando a sua exploração. De fato, é necessária a execução de muitas atividades para se realizar uma pesquisa mineral e nem sempre ela leva à descoberta de áreas exploráveis economicamente (SINGER and KOUDA, 1999a).

Segundo (PEREIRA, 2003), a pesquisa inicia com a delimitação de áreas dentro de uma região que, em um primeiro momento, parecem possuir interesse econômico. Essas áreas são, então, submetidas a estudos no intuito de verificar a presença de depósitos minerais. A seguir, podem-se realizar avaliações mais detalhadas dos depósitos considerados interessantes para estimar a quantidade e qualidade de seus recursos. Somente com essas informações associadas a informações de mercado é possível verificar a viabilidade da exploração do depósito.

Um depósito é um conjunto rochoso formado por rochas das quais é possível extrair minério separadas por rochas estéreis, as quais não possuem interesse econômico (MARANHÃO, 1989). Percebe-se que alguns tipos de depósitos possuem suas rochas mineralizadas separadas em maiores distâncias por áreas estéreis do que outros tipos. Percebe-se também que existe diferença entre o tamanho (área) médio de um tipo de depósito em relação a outros tipos. Futuros trabalhos de prospecção mineral possuem mais chances de sucesso se fizerem uso desses e de outros conhecimentos adquiridos com a experiência em exploração mineral. Uma maneira de integrar essas informações é através de modelos de depósitos minerais (SINGER, 2006).

2.2 Modelos de Depósitos Minerais

Modelos de depósitos minerais constituem uma importante ferramenta para integrar informações como a geologia do depósito, análises geoquímica e geofísica, história da exploração, depósitos e ocorrências conhecidas (SINGER, 1996a). Eles são elaborados com o objetivo de representar conhecimento útil a diferentes atividades na exploração mineral. A Figura 2.1 apresenta esquematicamente o papel desempenhado por modelos de depósito mineral.

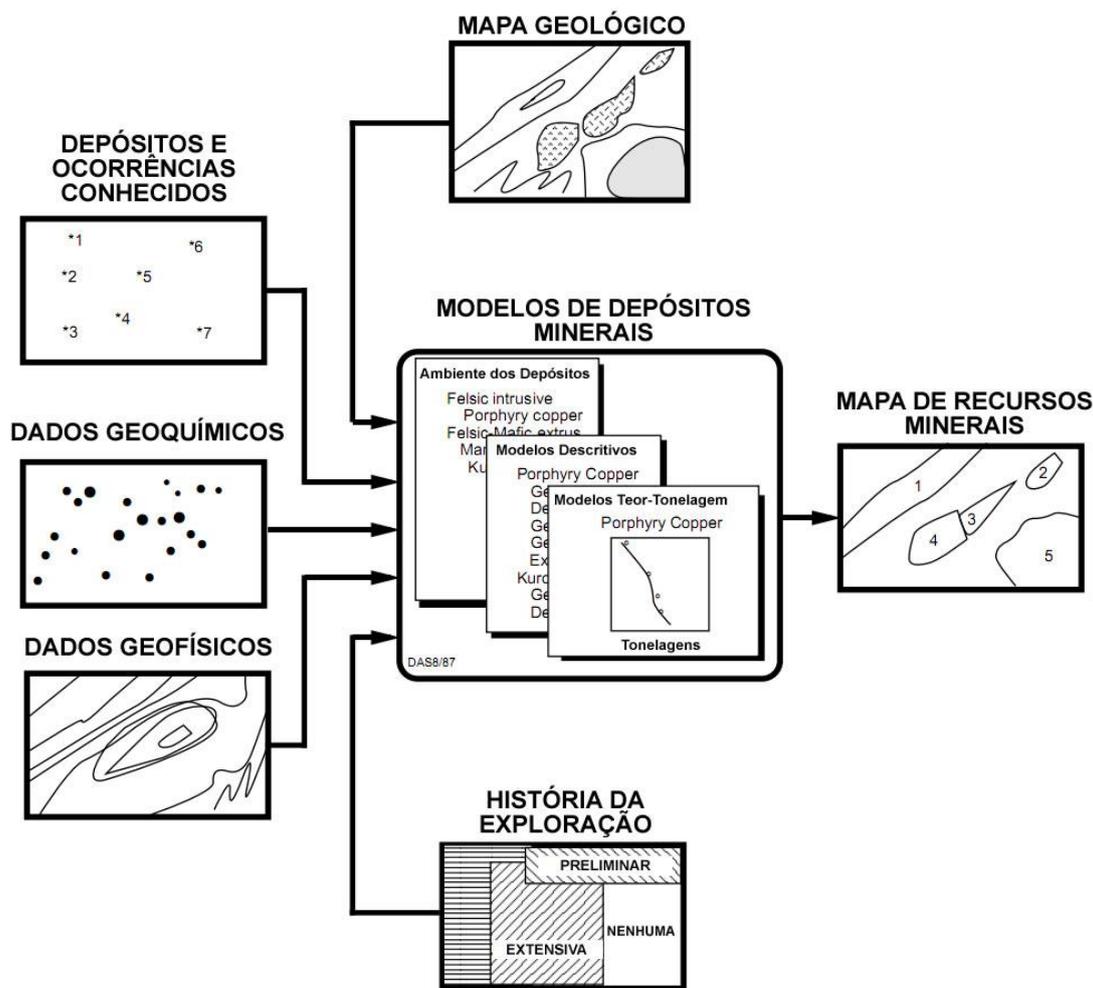


Figura 2.1 – A organização de informações em modelos de depósitos minerais. Fonte: Adaptada de (SINGER, 1996a).

Modelos de depósitos minerais existem, portanto, para auxiliar a busca por respostas a questionamentos específicos da pesquisa mineral. Esses questionamentos podem ser relativos, por exemplo, à fase de exploração mineral e auxiliarem no processo de delimitação de áreas de estudo prioritárias. De outro modo, podem ser relativos à fase de avaliação de depósitos e auxiliarem na geração de estimativas sobre os depósitos. Podem, enfim, serem elaborados para auxiliar em diversas atividades de qualquer das três fases da prospecção mineral. Para isso, é importante que representem o conhecimento de maneira que possa ser aplicado a novos estudos. Isto significa que devem ser elaborados com base em dados confiáveis e um conjunto de regras deve ser válido para todos os dados utilizados (AMORIM *et al*, 2005).

A avaliação quantitativa de recursos minerais utiliza modelos quantitativos para prover informações úteis para análises de custos e riscos de investimento em atividades de prospecção mineral. Esses modelos podem ser, por exemplo, modelos descritivos, modelos teor-tonelagem e modelos de densidade de depósito (SINGER, 2008a).

Modelos descritivos são formados de duas partes (COX and SINGER, 1986): a primeira identifica o ambiente geológico e configurações tectônicas das regiões onde pode ocorrer um tipo de depósito mineral, a segunda parte identifica características essenciais dos depósitos para esse tipo. Estes modelos são úteis na delimitação das áreas que possibilitam a ocorrência de depósitos de um determinado tipo dentro da região analisada. Assim, sua organização procura evidenciar informações litológicas das rochas e configurações tectônicas dos depósitos de um determinado tipo por serem essas informações facilmente obtidas de mapas geológicos.

Os modelos teor-tonelagem possuem informações sobre as tonelagens e os teores médios de alguns minerais dos depósitos (COX and SINGER, 1986). A tonelagem se refere à quantidade total de recursos do depósito no menor ponto de corte de teor possível. Os teores são relativos aos minerais de possível interesse econômico encontrados nos depósitos. Esses modelos possuem: a descrição geológica do ambiente dos depósitos que o compõe, a lista de depósitos com seus respectivos quantitativos de teores e tonelagem e representações gráficas que mostram os teores ou tonelagens no eixo horizontal e a proporção acumulada de depósitos no eixo vertical.

Modelos teor-tonelagem confiáveis são baseados em depósitos bem conhecidos de um determinado tipo e todos os depósitos seguem um conjunto específico de regras. Uma das regras é que o ambiente geológico em que ocorrem seja consistente com o modelo descritivo de um determinado tipo. Uma outra regra poderia determinar que depósitos separados em menos de dois quilômetros por uma área estéril sejam considerados como um único depósito. Respeitadas as regras de consistência interna do modelo, as informações quantitativas presentes nos modelos teor-tonelagem podem ser utilizadas para auxiliar atividades importantes da prospecção mineral. Uma utilidade aplicada em várias pesquisas é a estimativa de teores e tonelagens de depósitos não descobertos em ambientes geológicos semelhantes ao de um modelo publicado (ZHAO, 2002; YAN *et al*, 2007; SINGER, 1996a).

Modelos de densidade de depósito são particularmente úteis na geração de estimativas sobre o número de depósitos não descobertos. Neste sentido, modelos de densidade de depósito são úteis para guiar as estimativas do número de depósitos não descobertos do mesmo modo que as frequências dos teores e tonelagens são modelos do tamanho e qualidade de depósitos não descobertos (SINGER, 2008b).

As densidades representam o número de depósitos por uma medida de área que são adquiridos a partir de áreas de controle bem exploradas de diversas partes do mundo. Um histograma pode ser construído para evidenciar a frequência em que determinada densidade de depósito ocorre. A Figura 2.2 apresenta uma situação hipotética na qual as frequências das densidades de um tipo de depósito relativas a 12 áreas bem exploradas são representadas através de um histograma.

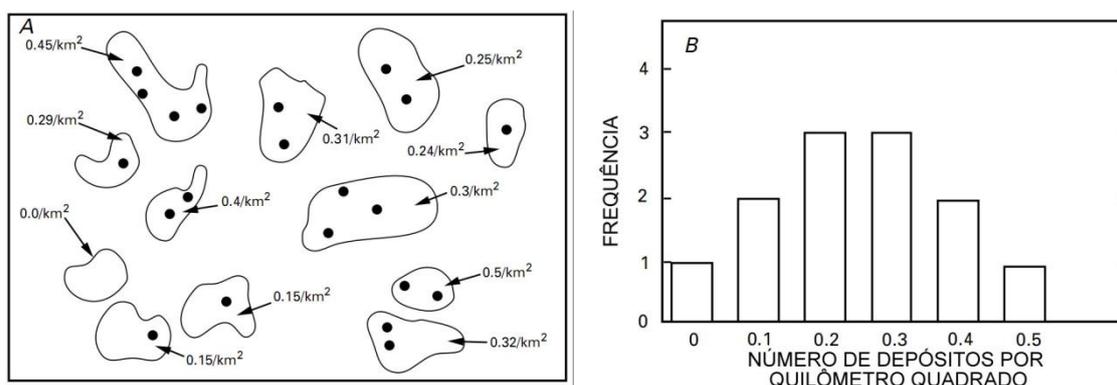


Figura 2.2 – Representação de uma situação hipotética para construção de modelo de densidade de depósitos: (a) 12 regiões bem exploradas com suas respectivas densidades e (b) histograma das densidades de depósito derivadas. Fonte: Adaptada de (SINGER *et al*, 2001).

Estimativas sobre a densidade devem considerar um tipo particular de depósito dentro de uma região delineada em que as configurações geológicas permitam a existência desse tipo de depósito e, ainda, deve existir um modelo teor-tonelagem para esse mesmo tipo de depósito. Durante o desenvolvimento de modelos de densidade de depósito é importante que as mesmas regras aplicadas em modelos descritivos ou teor-tonelagem sejam também aplicadas às densidades dos depósitos (SINGER, 2008b).

Os modelos de densidade de depósito são projetados para serem utilizados dentro do sistema de avaliação em três partes, o qual exige grande consistência (tanto a consistência interna dos modelos, quanto entre os modelos utilizados) para conseguir um grau de confiança aceitável na verificação de recursos de uma região. Somente quando utilizado nessa sistemática, o modelo é capaz de gerar estimativas com valor para a análise de recursos.

2.3 Avaliação Quantitativa de Recursos Minerais

Avaliações sobre recursos minerais acessíveis são importantes em estudos que visam fundamentar o planejamento do desenvolvimento econômico, as decisões sobre a utilização da terra, o compartilhamento de recursos minerais e as políticas relacionadas ao desenvolvimento sustentável (YAN, 2007; HAMMARSTROM, 2006). Avaliações quantitativas são particularmente interessantes quando conseguem explicitar as incertezas associadas às avaliações e suas partes constituintes e, dessa forma, servir como base para análises econômicas. Por essa razão, as estimativas resultantes serão melhores aceitas se a sua produção minimizar a possibilidade de influências que as tornem tendenciosas.

O sistema de avaliação em três partes apresenta uma metodologia que propicia a geração de estatísticas sobre depósitos não descobertos com um mínimo de influência que torne os resultados tendenciosos. As três partes são: (1) a delimitação de áreas em que a geologia permite a ocorrência de determinados tipos de depósitos; (2) a geração de estimativas sobre o número de depósitos não descobertos nas áreas para cada tipo de depósito; (3) a geração de estimativas sobre as quantidades de minério e sobre algumas características dos minerais dessas áreas a partir de modelos teor-tonelagem. A Figura 2.3 apresenta esquematicamente a avaliação quantitativa de recursos minerais em três partes em sua aplicação ao estudo de regiões do Estado de Nevada.

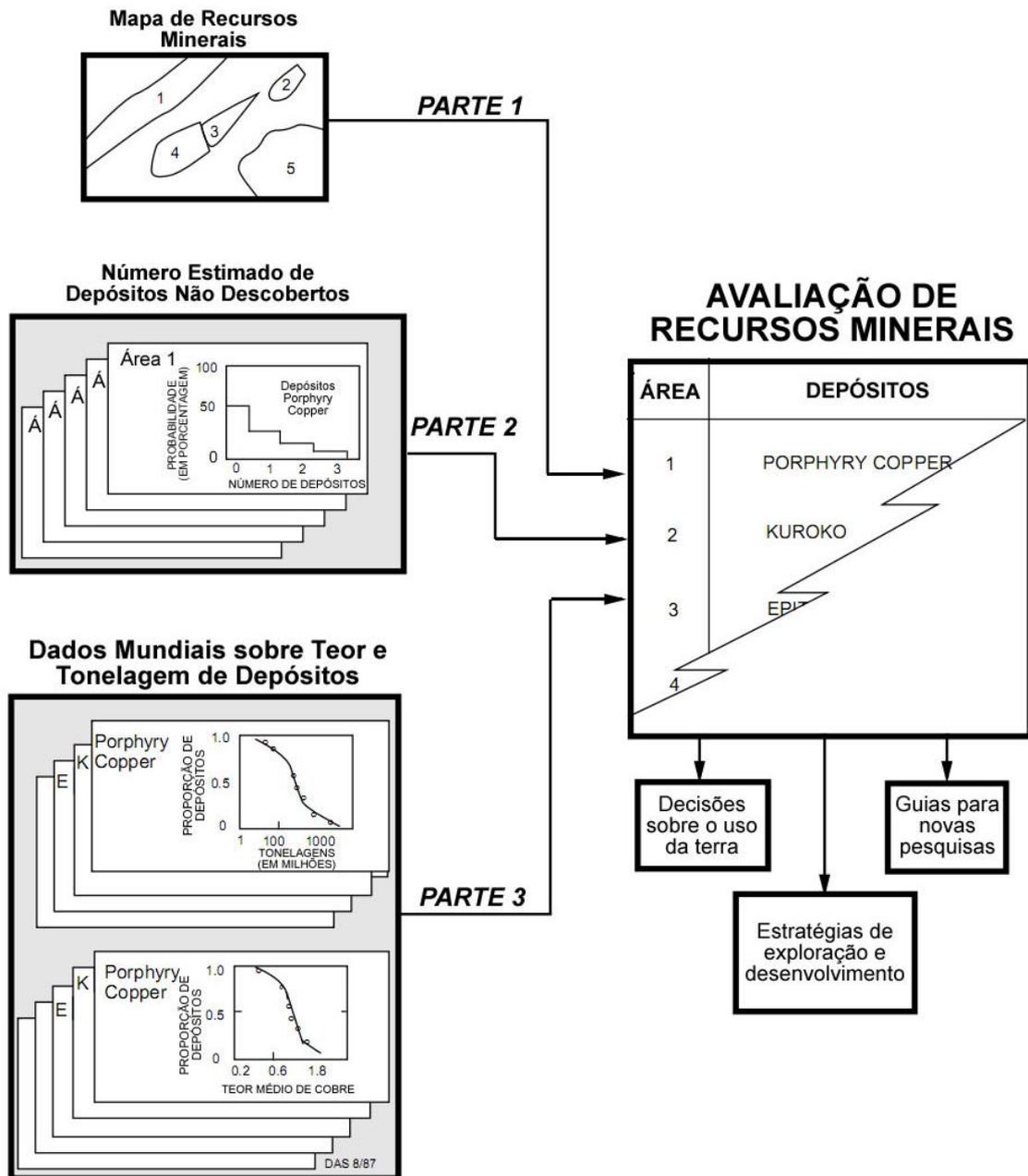


Figura 2.3 – Esquema do sistema de avaliação em três partes. Fonte: Adaptada de (SINGER, 1996a).

Na primeira parte, com base nas configurações geológicas da região em estudo e na descrição da geologia permissiva de cada tipo de depósito mineral, são delineadas as áreas permissivas para um ou mais tipos de depósito. As descrições relativas a cada tipo são encontradas em modelos descritivos de depósitos minerais elaborados a partir de depósitos da região ou, como ocorre com maior frequência, de depósitos conhecidos de diferentes partes do mundo que ocorrem em ambientes semelhantes ao da região em

estudo e possuem características similares aos depósitos e ocorrências conhecidos da região. Além disso, nessa parte, são considerados os tipos resultantes da classificação de depósitos conhecidos da região para refinar a delimitação das áreas permissivas.

Na segunda parte um número é associado à estimativa da quantidade de depósitos de cada tipo nas áreas delimitadas. O sistema de avaliação em três partes não define métodos específicos para gerar as estimativas. No entanto, pode-se destacar o uso de modelos de densidade de depósitos em regiões onde não há muitas informações, o que o torna bastante útil para prever o número de depósitos não descobertos em regiões desconhecidas. Nesse caso, é preciso que as densidades sejam calculadas para um tipo de depósito em particular, considere somente a área onde a geologia permite a existência desse tipo de depósito e que exista um modelo teor-tonelagem para o tipo considerado.

A aplicação de modelos de densidade de depósito se baseia em analogia entre áreas bem exploradas, onde se conhece as densidades dos depósitos por unidade de área, e a área da região em estudo. Um exemplo é a sua utilização para prever o número de depósitos não descobertos do tipo *Low-sulfide Au-quartz vein* na Venezuela (COX, 1993). Ao final, as estimativas geradas nesta segunda etapa devem ser consistentes com o modelo teor-tonelagem identificado para a região.

Na terceira etapa, modelos teor-tonelagem consistentes com os modelos descritivos são utilizados para estimar os teores e tonelagens dos depósitos da região.

Em (SINGER, 1993) afirma-se que o sistema de avaliação em três partes consegue produzir estimativas válidas porque a sistemática adotada segue as seguintes regras de consistência: as áreas delimitadas são consistentes com os modelos descritivos, os modelos teor-tonelagem são consistentes com os modelos descritivos, os modelos teor-tonelagem são consistentes com os depósitos conhecidos da área e as estimativas do número de depósitos são consistentes com os modelos teor-tonelagem. A seleção de modelos apropriados é inicialmente guiada pela determinação do tipo de depósitos da região, o que torna a classificação de depósitos um ponto importante para a correta aplicação dessa metodologia.

2.4 Classificação de Depósitos Minerais

A aprendizagem de máquina é uma sub-área da Inteligência Artificial que se dedica ao desenvolvimento de algoritmos e técnicas que permitam ao computador adquirir e organizar conhecimentos para aperfeiçoar seu desempenho na execução de alguma tarefa (MONARD, 1997). A atuação da aprendizagem de máquina se estende de problemas simples a complexos em relação à natureza do domínio e dos dados que deve manipular. Como exemplo pode-se citar estudos de comportamento social (AMORIM *et al.*, 2008), aplicações industriais (COSTA, 2006) e ciências naturais (DE'ATH and FABRICIUS, 2000).

Uma forma de representar conhecimento é através de um conjunto de regras propostas por um especialista no domínio. Neste caso, a resolução de um problema se apóia na aplicação de regras manipuláveis pelo computador que representam o conhecimento (experiência) dos especialistas do domínio. Esse modelo de representação é aplicável quando se conhece a natureza dos dados que serão manipulados e as regras inerentes ao domínio do problema, mas torna a resolução de problemas dependente do conhecimento de especialistas e exclui problemas de natureza complexa.

Em um sentido oposto, considera-se a possibilidade de extrair o conhecimento necessário à resolução de problemas a partir de um conjunto de dados que são apresentados como casos específicos do problema. Nesse ponto faz sentido a busca por técnicas que permitem a aquisição de conhecimentos a partir de dados. No aprendizado por indução, um conceito é adquirido através de inferências indutivas sobre os casos específicos apresentados e, desse modo, os modelos gerados podem ou não preservar a verdade (COSTA, 2006). Em outras palavras, pode-se dizer que as conclusões conseguidas com o raciocínio indutivo são estatisticamente corroboradas pela verificação dos casos utilizados no aprendizado.

O aprendizado indutivo pode ser classificado em aprendizado supervisionado e aprendizado não supervisionado. O conhecimento prévio necessário à indução da solução para novas situações provêm de um conjunto de casos ou situações previamente conhecidas em que todos os casos são descritos pelo mesmo conjunto de características. Se, juntamente com as características dos casos, são também fornecidas as saídas corretas que lhe correspondem, o objetivo do aprendizado se realiza na construção de

um modelo que tenta ajustar as características dos casos apresentados à saída esperada. Isto configura um problema manipulável pelo aprendizado supervisionado. No entanto, se não é fornecida qualquer saída e a tentativa é de explorar os dados existentes para extrair conhecimento novo a partir deles, temos um problema adequado ao aprendizado não-supervisionado.

Dito de outra forma, o aprendizado supervisionado é adequado a problemas classificatórios (classificação e regressão) nos quais se dispõe de um conjunto de características e de rótulos (as classes ou saídas correspondentes) das situações que representam o conhecimento prévio do problema. Nesse caso, técnicas de aprendizado de máquina supervisionado são utilizadas para treinamento de algoritmos através de um conjunto de dados cujas características e classes são conhecidas na intenção de gerar um classificador que permita induzir, com aceitável nível de confiança, uma classe para dados que tem somente as suas características conhecidas.

Na prática, ao verificar a aplicabilidade de técnicas de aprendizado de máquina, fala-se em conjunto de treinamento e conjunto de testes. Em ambos os conjuntos, conhece-se as características e as classes dos dados, mas o conjunto de treinamento é utilizado para a construção do modelo indutivo (classificador) e o conjunto de teste é aplicado posteriormente para verificar a aplicabilidade do classificador através de uma medida de desempenho. Uma medida de desempenho comum é a taxa de acerto, ou seja, a razão entre o número de casos de teste classificados corretamente e o número de casos utilizados para teste. Após a verificação do classificador mais adequado ao problema, o modelo indutivo gerado pode então ser aplicado para classificar situações novas (casos que somente se conhece as características) dentro da medida de confiabilidade oferecida pelo classificador.

A classificação de depósitos minerais no contexto desse trabalho situa-se no domínio da prospecção mineral, mais precisamente no auxílio à avaliação quantitativa de depósitos não descobertos de uma região. O sistema de avaliação em três partes mostra-se adequado nesse contexto e, em razão de sua utilização, recorre-se à classificação de depósitos com o objetivo de identificar os modelos de depósitos mais adequados em cada parte ou etapa da sistemática. A consistência dessa sistemática está fortemente relacionada ao modelo teor-tonelagem, como pode ser visto em (SINGER e KOUDA, 1988; SINGER, 1996b). Assim é conveniente que a classificação de depósitos

da região vise identificar tipos relacionados a modelos teor-tonelagem. Os cerca de 60 modelos teor-tonelagem construídos em (COX and SINGER, 1986) têm sido aplicados como modelos gerais na aplicação do sistema de avaliação em três partes, por isso as classes de depósitos podem ser tipos relacionados a esses modelos.

Em (SINGER, 2006) afirma-se que os teores e tonelagens da maioria dos tipos são significativamente diferentes para cada tipo de depósito. E, ainda, que dados sobre teores e tonelagens são um tipo de informação que normalmente está acessível nos estágios iniciais do desenvolvimento da prospecção e também para depósitos em localizações remotas. Além disso, considera a melhoria de resultados pela identificação da presença de algumas rochas hospedeiras. As características dos depósitos utilizadas para este problema de classificação podem, portanto, ser os dados sobre tonelagem e teores minerais acessíveis dos depósitos e a identificação da presença de tipos de rochas hospedeiras.

O problema de identificar o modelo teor-tonelagem configura-se, assim, como o problema de classificar corretamente os depósitos da região em estudo em tipos, em que cada tipo possui um modelo teor-tonelagem associado. Assim, as classes dos depósitos são tipos de depósitos que possuem um modelo teor-tonelagem associado e as características estão relacionadas à descrição de depósitos nesses modelos.

A respeito da possibilidade quase certa de que os modelos gerados a partir dos dados não ofereçam 100% de certeza na indução de resultados, é importante ressaltar que diferentes técnicas ou algoritmos de aprendizagem comportam-se diferentemente e, assim, apresentam diferentes resultados quando aplicados a um domínio. A variação do domínio e dos casos específicos disponíveis ao aprendizado também provocam a divergência de resultados de maneira que não se pode afirmar qual a técnica mais adequada à resolução de um determinado problema. Assim, para verificação da possibilidade de aplicar técnicas de aprendizado de máquina a um novo domínio é interessante que se realize a investigação da adequabilidade de diferentes técnicas. Neste trabalho, realizou-se a investigação de três técnicas que representam diferentes paradigmas de aprendizado de máquina.

A técnica Árvores de Decisão representa o paradigma de aprendizado lógico ou simbólico. Sistemas de aprendizado simbólico aprendem conceitos geralmente compreensíveis por seres humanos que podem ser representados na forma de

hiperplanos (MILARÉ, 2003). A possibilidade de compreensão dos conceitos ou regras utilizadas torna técnicas de aprendizado simbólico interessantes.

O paradigma conexionista engloba técnicas que utilizam unidades altamente conectadas. As Redes Percéptron Multicamadas são representantes desse paradigma. A capacidade de resolver problemas complexos e não-lineares explica sua aplicação com sucesso em vários domínios.

Redes Neurais Probabilísticas representam o paradigma baseado em exemplos (do inglês *instance-based*). Os algoritmos que seguem esse paradigma lentamente aprendem a partir de exemplos ou casos e baseiam suas decisões para novas situações na similaridade destas com os casos que lhe são conhecidos. Em geral o treinamento do algoritmo é mais rápido que o treinamento de outros tipos de redes neurais, porém a velocidade de execução diminui em razão da quantidade de processamento necessário para verificar a similaridade de uma nova situação a todos os casos do modelo.

2.5 Árvores de Decisão

Árvores de Decisão (AD) são algoritmos de aprendizado simbólico que sintetizam conceitos relativos a um conjunto de casos através de regras resultantes do particionamento recursivo sobre os atributos dos casos (COSTA, 2006; MONARD, 1997). Isso significa que é possível explicar como a decisão é feita ao seguir a estrutura da árvore que se forma pelas sucessivas regras sobre os atributos dos casos. Árvores usadas para resolver problemas de classificação são denominadas árvores de classificação (SOBRAL, 2003). Uma árvore de classificação é construída a partir de um conjunto de casos, chamado conjunto de treinamento, previamente classificados que podem ser descritos através de atributos contínuos ou discretos. Supõe-se que cada caso ou exemplo utilizado para construção da árvore possua valores diferentes em ao menos um de seus atributos. Então esse conjunto é dividido com base em uma expressão lógica formulada sobre valores dos atributos. A expressão poderia ser, por exemplo, a verificação se um atributo é menor que um determinado valor, para o caso de atributos contínuos, ou a verificação se um atributo pertence a um subconjunto de valores, para atributos discretos.

O resultado do particionamento são subconjuntos mutuamente exclusivos cuja união contém todos os casos do conjunto inicial. A recursividade do algoritmo é caracterizada em razão de cada subconjunto formado ser tratado de maneira similar ao conjunto inicial e, dessa forma, sucederem novas divisões dos subconjuntos até que todos os elementos do subconjunto pertençam a uma mesma classe ou um critério de parada se torne verdadeiro. Se toda divisão resultar em dois subconjuntos, diz-se que a árvore é binária. Árvores binárias são mais frequentemente usadas devido a sua simplicidade e ao conhecimento extensivo de suas características (COSTA, 2006).

O efeito das sucessivas divisões sobre o espaço formado pelos atributos é a formação de hiper-retângulos de forma que a interseção destes é vazia e a união é todo o espaço (BERK, 2006). A Figura 2.4 esquematiza uma árvore de decisão binária construída a partir de 2 características predictoras ou atributos e sua representação no espaço de atributos.

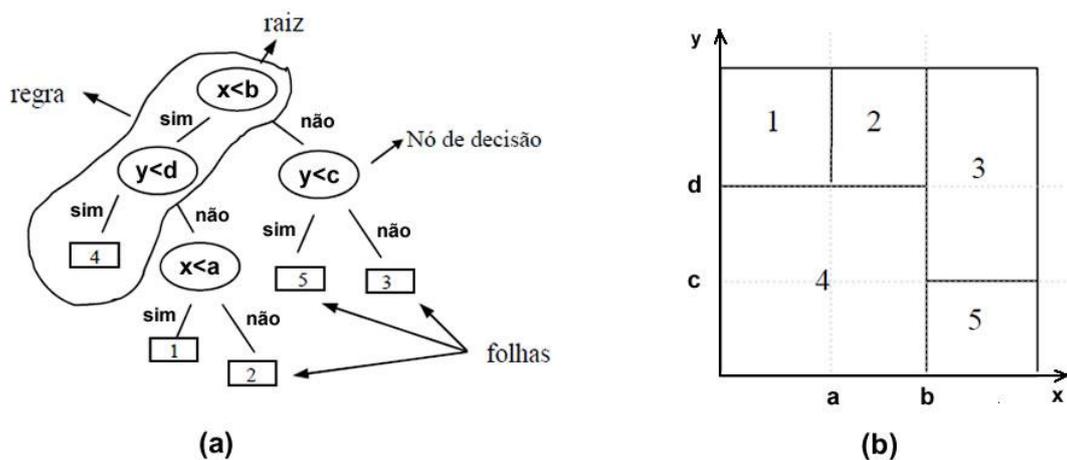


Figura 2.4 – Representação de uma árvore de decisão binária: (a) estrutura da árvore e (b) sua respectiva representação no espaço de atributos. Fonte: Adaptada de (GAMA, 2002).

A estrutura de uma árvore consiste de uma hierarquia de nós internos e externos que estão conectados através de ramos e cuja disposição pode ser representada como uma árvore invertida (SOBRAL, 2003). Cada nó interno, chamado nó decisório, representa uma unidade de tomada de decisão que avalia qual será o próximo nó descendente. Cada possibilidade de resultado da avaliação é representada por um ramo da árvore. E cada nó externo, também conhecido como folha, representa um rótulo ou

classe. O primeiro nó decisório é a raiz da árvore. O caminho que leva a cada folha, a partir da raiz, é único e corresponde a uma regra de decisão formada pela aplicação do conectivo lógico AND a todos os testes que ocorrem neste caminho.

No espaço de atributos, cada nó decisório é representado por uma linha perpendicular ao eixo correspondente ao atributo de decisão a qual é traçada a partir do valor considerado na verificação realizada naquele nó. Assim, a raiz da árvore, que contém o teste $x < b$, é responsável pela primeira divisão no espaço, a qual é realizada pela linha perpendicular ao eixo x que o toca no valor b . De maneira similar, os demais nós internos da árvore correspondem às demais divisões do espaço representado na Figura 2.4b. Ao final, o espaço encontra-se dividido em regiões retangulares, cada qual correspondendo a um rótulo. Isso significa que todo caso cujos atributos o localizarem dentro de uma região será classificado com o rótulo correspondente àquela região.

Um dos mais conhecidos algoritmos para construção de árvores de decisão é o CART – *Classification and Regression Trees* (BREIMAN *et al*, 1984). A árvore construída pelo CART resulta de duas etapas: o crescimento extensivo da árvore e a poda da árvore.

2.5.1 Crescimento da Árvore

O crescimento de uma árvore ocorre através de sucessivas divisões sobre o conjunto de treinamento. As divisões são baseadas em critérios que procuram maximizar a homogeneidade dos atributos dos casos pertencentes a cada subconjunto da divisão, o que, para árvores de classificação, significa que a maior parte dos elementos de um subconjunto deve ser de uma única classe. Os elementos do subconjunto que não são da classe a que ele está associado evidenciam a existência de erro na árvore. Dessa forma, os critérios utilizados para selecionar a melhor divisão em cada momento devem considerar o erro resultante, o que normalmente é calculado através de uma função de impureza.

Para melhor explicar o significado de impureza, supõe-se que uma árvore de decisão tenha dois resultados possíveis: A ou B. Numa situação ideal, para um nó

terminal, todos os casos que compõem esse nó deveriam ser da classe A ou todos deveriam ser da classe B e, então, a sua impureza seria a menor possível. No entanto, se metade dos casos fossem da classe A e a outra metade fosse da classe B, o nó seria o mais impuro possível.

A função que determina a impureza de um nó para problemas de classificação é uma função das probabilidades de um elemento ser de uma classe, dado que ele pertence ao conjunto de elementos daquele nó. Ou seja, a função de impureza ϕ mapeia os argumentos de entrada p_1, p_2, \dots, p_k , com $\sum_{i=1}^k p_i = 1$, tal que as seguintes condições são satisfeitas:

- $\phi(1/k, 1/k, \dots, 1/k) = \text{máximo}$
- $\phi(1, 0, 0, \dots, 0) = \phi(0, 1, 0, \dots, 0) = \phi(0, 0, 0, \dots, 1) = 0$

em que $p_i, \forall i \in [1..k]$ é a porcentagem de elementos da classe i no nó, k é o número de classes possíveis e $\phi(p_1, p_2, \dots, p_k)$ é a função de impureza. Portanto, a função de impureza resulta em valor máximo quando todas as classes estão misturadas no nó e resulta em 0 (zero) quando o nó contém casos de uma única classe.

Entre as funções de impureza mais conhecidas estão a entropia, na Equação 2.1, e o índice de Gini, na Equação 2.2 (BERK, 2006).

$$\phi_e(p_1, \dots, p_k) = -\sum_{i=1}^k p_i \log(p_i) \quad \text{Eq. 2.1}$$

$$\phi_g(p_1, \dots, p_k) = 1 - \sum_{i=1}^k p_i^2 \quad \text{Eq. 2.2}$$

Essas funções comportam-se de maneira diferente quando a árvore possui mais de duas classes possíveis como resposta. O índice de Gini tende a isolar em um ramo casos que representam uma classe com quantidade relativamente pequena de representantes. Já a entropia balanceia o número de casos em cada ramo, ou seja, tende a selecionar divisões que são similares em tamanho e homogeneidade. Assim, o índice de Gini é mais recomendado para esses casos (BREIMAN *et al*, 1984).

A qualidade de uma divisão s para um nó τ que resulta nos nós τ_e e τ_d pode ser formalmente representada como:

$$\Delta i(s, \tau) = i(\tau) - p(\tau_e)i(\tau_e) - p(\tau_d)i(\tau_d) \quad \text{Eq. 2.3}$$

onde $i(\tau)$, $i(\tau_e)$ e $i(\tau_d)$ são, respectivamente, os valores da impureza no nó pai, no nó descendente da esquerda e no nó descendente da direita e $p(\tau_e)$ e $p(\tau_d)$ são, respectivamente, as probabilidades de estar no nó da esquerda e no nó da direita.

O algoritmo CART calcula $\Delta i(s, \tau)$ para todas as divisões em cada atributo e escolhe o atributo e a divisão neste atributo que possui maior valor. Uma divisão sobre um atributo contínuo ou ordenado deve selecionar um valor para dividir a amostra através de expressões do tipo $x < b$. Para atributos categóricos deve-se selecionar um subconjunto S de valores para dividir a amostra através de expressões do tipo $x \in S$.

O mesmo procedimento é aplicado aos nós descendentes e, dessa forma, a árvore pode crescer continuamente até que todos os elementos do conjunto de treinamento estejam perfeitamente classificados. A árvore resultante é a árvore de tamanho máximo. Se o conjunto usado para construir a árvore possui ruídos ou não é representativo, o que é muito comum, a árvore resultante não é ideal para classificar elementos que não pertencem ao conjunto de treinamento. Nesse caso fala-se que há *overfitting*.

O *overfitting* de uma árvore de decisão T é caracterizado pela existência de uma árvore T' que apresenta menor erro do que T quando aplicada ao conjunto total de elementos, embora T possua menor erro do que T' quando aplicada ao conjunto de treinamento (KOTSIANTIS, 2007). Segundo (BREIMAN *et al*, 1984), para prevenir o problema de *overfitting* deve-se realizar a poda da árvore.

2.5.2 Poda da Árvore

A poda da árvore visa melhorar a capacidade de classificação pela prevenção da super-especialização da árvore no conjunto de treinamento, o que caracteriza *overfitting*. Duas abordagens podem ser utilizadas: parar o algoritmo de treinamento antes que ele alcance a perfeita classificação dos dados de treino, conhecida como pré-poda, ou podar a árvore induzida, conhecida como pós-poda (KOTSIANTIS, 2007).

O modo mais direto de prevenir *overfitting* é a pré-poda de forma a não permitir que a árvore cresça até seu tamanho máximo. Isso é feito através do estabelecimento de um critério que, quando se torna verdadeiro, impede a realização de novas divisões. Esse critério poderia ser um número mínimo de nós que, se um nó descendente não contiver, impossibilita a ocorrência da divisão. De outra forma, pode-se estabelecer um critério não trivial como um teste de limite para uma métrica de qualidade. É possível que a pré-poda selecione uma árvore subótima ao interromper o crescimento da árvore (BREIMAN *et al*, 1984).

O procedimento de pós-poda ocorre após a geração da árvore de tamanho máximo. Sub-árvores são geradas pela remoção de nós internos e associação destes às classes que mais ocorrem entre os casos de treino que compõem aqueles nós. Algumas sub-árvores encontradas nesse processo são avaliadas e a melhor delas é escolhida.

Supõe-se um critério de qualidade para uma árvore T definido como:

$$R(T) = \sum_{\tau \in \Psi} p(\tau)r(\tau) \quad \text{Eq. 2.4}$$

onde Ψ é o conjunto de nós terminais de T , $p(\tau)$ é a probabilidade de um caso cair no nó terminal τ e $r(\tau)$ é uma medida de qualidade do nó τ . Então o objetivo da poda é encontrar a sub-árvore T^* a partir da árvore de tamanho máximo T_{\max} que minimiza $R(T)$. No algoritmo CART o valor de $R(T)$ é uma função dos erros de classificação e da complexidade da árvore.

Para considerar a complexidade da árvore, define-se uma função objetivo que inclui um parâmetro de penalidade à complexidade associada ao número de nós terminais, como na Equação 2.5.

$$R(T) = R(T) + \alpha |\psi| \quad \text{Eq. 2.5}$$

onde $\alpha \geq 0$ é um parâmetro de complexidade e $|\psi|$ é o número de nós terminais da árvore T . Assim, o tamanho da árvore influencia o valor do critério de qualidade e, como o objetivo é minimizá-lo, se duas sub-árvores possuírem o mesmo valor de erro, aquela que possui menos nós terminais será a considerada a melhor sub-árvore.

Existe apenas uma sub-árvore que seja a menor e que minimize o custo de complexidade para um valor específico de α (BREIMAN *et al*, 1984). Dessa forma, a árvore resultante da pós-poda é a sub-árvore ótima.

O algoritmo CART primeiramente gera a árvore de tamanho máximo e, a partir dela, realiza pós-poda para selecionar a sub-árvore ótima que atuará como modelo de classificação dos casos.

2.6 Rede Percéptron Multicamadas

A Rede Percéptron Multicamadas (MLP, do inglês *Multilayer Perceptrons*) é uma Rede Neural Artificial altamente conectada capaz de resolver problemas complexos e não-lineares (FLAKE, 1998). Redes Neurais Artificiais (RNA) são máquinas de aprendizado distribuídas, adaptativas e geralmente não-lineares construídas a partir de várias unidades de processamento independentes (PRÍNCIPE, 2000). As unidades de processamento são conhecidas como neurônios artificiais. Um modelo de Rede Neural Artificial é constituído de neurônios interconectados, onde cada conexão possui um peso associado.

O modelo de neurônio artificial de McCullon e Pits é capaz de receber um número m de valores ou sinais de entrada e enviar um único valor de saída (HAYKIN, 2001). Inicialmente os sinais são processados através de um somatório das entradas associadas aos pesos de suas conexões de entrada e a um escalar associado ao neurônio conhecido como *bias*, tal como na Equação 2.6.

$$v_k = \sum_{j=i}^m w_{kj} x_j + b_k \quad \text{Eq. 2.6}$$

onde k é o número que identifica o neurônio, m é o número de sinais de entrada, x_j é o valor da entrada j , w_{kj} é o peso associado ao neurônio k para a entrada j e b_k é o valor *bias* associado ao neurônio k .

Os valores v_k são, em seguida, aplicados a uma função de ativação φ para se obter o valor de saída do neurônio, tal como na Equação 2.7.

$$y_k = \varphi(v_k) \quad \text{Eq. 2.7}$$

As funções de ativação normalmente utilizadas são (PRÍNCIPE, 2000):

- Função limiar: limita a saída a dois valores, tal como na Equação 2.8;

$$y_k = \begin{cases} +1, v_k \geq 0 \\ 0, v_k < 0 \end{cases} \quad \text{Eq. 2.8}$$

- Função tangente hiperbólica: função sigmóide que permite que a saída seja suavizada pelo cálculo da Equação 2.9;

$$y_k = \tanh(v_k) \quad \text{Eq. 2.9}$$

- Função logística: função sigmóide que permite que a saída seja suavizada pelo cálculo da Equação 2.10.

$$y_k = \frac{1}{1 + \exp(v_k)} \quad \text{Eq. 2.10}$$

Os neurônios de um modelo de Rede Neural Artificial são organizados em camadas de modo que dois neurônios quaisquer da RNA pertencem à mesma camada se o número mínimo de conexões necessárias para se chegar ao neurônio a partir de uma entrada da RNA é o mesmo. A primeira camada de uma rede, denominada camada de entrada é aquela que recebe os sinais de entrada da RNA. A última camada, denominada camada de saída, é aquela que oferece a saída final da RNA. Além dessas, entre uma e outra, podem existir outras camadas que não são visíveis fora da RNA, denominadas camadas ocultas.

Uma característica importante das RNAs é a maneira que os neurônios são conectados, conhecido como topologia da rede. Segundo essa característica, as redes podem ser divididas em: RNAs recorrentes e RNAs não-recorrentes.

Em RNAs recorrentes podem existir conexões entre neurônios da mesma camada. Além disso, existe retroalimentação, ou seja, o valor de saída de um neurônio pode alimentar o mesmo neurônio ou um neurônio que está em uma camada anterior a que ele pertence.

Em RNAs não-recorrentes não existe retroalimentação e todas as conexões ocorrem entre os neurônios de uma camada com os neurônios da camada imediatamente posterior. Redes com essa topologias são redes *feedforward*.

As Redes Percéptron Multicamadas são redes *feedforward* constituídas de uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída (Figura 2.5). A camada de entrada é formada por neurônios simples responsáveis pela distribuição dos dados de entrada para as camadas subsequentes. As camadas ocultas são constituídas por percéptrons, os quais são neurônios que recebem a soma ponderada dos sinais transmitidos pela camada anterior e aplicam uma função de ativação a esses dados, gerando uma saída para a camada seguinte. Normalmente os percéptrons de MLPs utilizam uma função sigmóide como função de ativação.

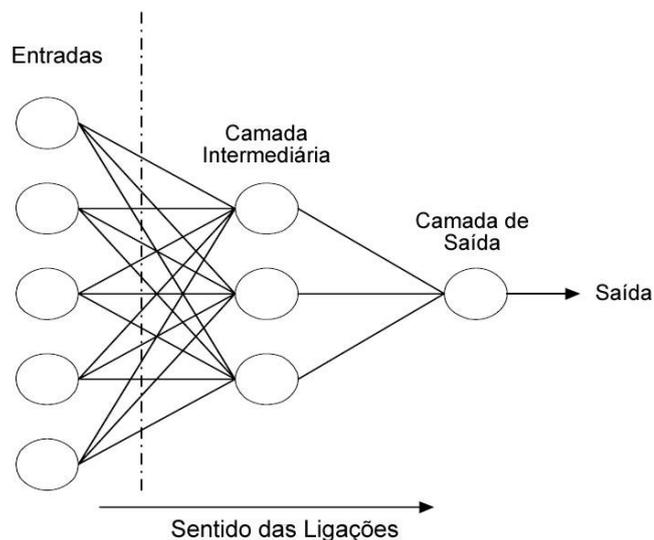


Figura 2.5 – Rede Percéptron Multicamadas. Fonte: (COSTA, 2006).

Além da identificação dos tipos de conexão que podem existir na rede, a definição do número de camadas ocultas da MLP é importante para definir o tipo de problema que ela pode tratar. Por exemplo, uma rede com apenas uma camada de neurônios de McCullon e Pits somente podem resolver problemas linearmente separáveis (COSTA, 2006). Segundo (PRÍNCIPE, 2000) uma MLP com duas camadas ocultas pode ser considerada um classificador universal, isto é, pode realizar qualquer mapeamento entrada-saída.

Outra decisão importante na construção de uma MLP é o número de neurônios de cada camada, pois se forem selecionados poucos neurônios a rede pode não ser capaz de se adequar a modelos de dados complexos. Em sentido oposto, muitos neurônios podem ser a causa de *overfitting* da rede. O número de neurônios na camada de entrada é igual ao número de sinais de entrada e o número de neurônios na camada de saída em uma MLP utilizada para classificação é o número de classes possíveis. Dessa forma, o problema é a identificação do número de unidades de processamento das camadas ocultas. Estudos sobre o *overfitting* de RNAs e sobre o número mínimo de neurônio e a quantidade de neurônios necessária para tratar um determinado problema são apresentados em (CAMARGO and YONEYAMA, 2001) e (KON and PLASKOTA, 2000).

A transmissão de dados entre as camadas é ponderada através de pesos, os quais são ajustados durante as sucessivas iterações do treinamento da rede. Inicialmente, os pesos são atribuídos de maneira randômica à rede. Em seguida os elementos do conjunto de treinamento são utilizados para ajuste dos pesos através da comparação entre o resultado esperado para o elemento e o resultado apresentado pela rede e modificação dos valores dos pesos no caso de ocorrência de erro (KOTSIANTIS, 2007).

Um conhecido algoritmo de treinamento que pode ser utilizado para treinamento de MLPs é o algoritmo de retropropagação (do inglês, *back-propagation*), cuja atuação assemelha-se ao algoritmo conjugado descendente (HAYKIN, 2001).

2.7 Rede Neural Probabilística

Rede Neural Probabilística (PNN, do inglês *Probabilistic Neural Network*) é um tipo de Rede Neural Artificial baseada no classificador Bayesiano capaz de aprender mapeamentos não-lineares e classificar elementos baseados em critérios de similaridade (NOEL, 2008). Trata-se de um exemplo de rede neural supervisionada, o que significa que um conjunto de classificações corretas (conjunto de treinamento) é fornecido à rede para que ela seja construída. Diferente de outras Redes Neurais, as quais são treinadas através do ajuste dos pesos nas ligações entre os neurônios de uma camada com os de outra camada, na PNN não existe essa necessidade. O processo de treinamento da PNN consiste no ajuste de parâmetros (no caso, os parâmetros sigma) intrínsecos à sua

definição e não no ajuste de pesos (MONTANA, 1992). Dessa forma, o processo de treinamento das PNNs é mais rápido, o que representa uma vantagem à sua utilização em relação às demais Redes Neurais.

Outro fato intrínseco à definição de PNN é o fato de que ela considera todos os elementos da amostra de treinamento na construção da rede, de acordo com sua arquitetura. Tal fato implica em grande consumo de memória durante sua execução, o que pode torná-la mais lenta (NOEL, 2008). Abordagens de agrupamento podem ser utilizadas para melhorar o desempenho da PNN através do agrupamento de elementos da amostra de treinamento que sejam similares, o que resulta na diminuição do tamanho da rede.

Para melhor entender a PNN e seu funcionamento é importante compreender: (1) a base teórica que a fundamenta, (2) sua arquitetura e (3) o fluxo seguido para realizar a classificação.

2.7.1 Base Teórica

As PNNs foram inspiradas na regra de classificação de Bayes e na teoria da estimação de funções de densidade de probabilidade (f.d.p.) de Parzen (GEORGIUO *et al*, 2004). O teorema de Bayes, bastante conhecido da estatística, fornece um método para efetuar classificações ótimas de forma que o classificador Bayesiano é usado como padrão para a avaliação de outros métodos (BERRAR *et al*, 2003). A proposta desse teorema é a minimização dos custos associados a más classificações de forma que, dado um elemento X e sua classificação preferencial na classe i , a condição presente na Equação 2.11 deve ser satisfeita.

$$h_i c_i f_i(X) > h_j c_j f_j(X) \quad \text{Eq. 2.11}$$

onde, para $\forall i, j \in [1, 2, \dots, k]$ em que k é o número de classes possíveis, temos que h_i (h_j) é a probabilidade *a priori* de o elemento pertencer à classe i (j), c_i (c_j) representa o custo da má classificação de um elemento pertencente à classe i (j) e f_i (f_j) é a função densidade de probabilidade da classe i (j).

Quando os valores são desconhecidos é possível assumir que a probabilidade *a priori* e os custos associados a uma classe são iguais para todas elas, caso em que a definição das f.d.p. será crítica para a correta classificação final (FALCÃO, 2002).

Um método para estimar funções de densidades de probabilidade associadas a uma amostra aleatória foi desenvolvido por Parzen em 1962 (MASTERS, 1995). Por meio dele redes neurais que aplicam a regra de Bayes podem ser construídas. O método consiste na representação de cada elemento da amostra de treinamento por uma função designada de núcleo de Parzen. Normalmente um núcleo gaussiano é utilizado e a f.d.p. pode ser calculada através da Equação 2.12 (MASTERS, 1995).

$$g(x) = \frac{1}{n} \sum_{i=1}^n e^{-\sum_{j=1}^m \left(\frac{x_j - x_{ij}}{\sigma_j} \right)^2} \quad \text{Eq. 2.12}$$

onde, n é o número de elementos da amostra, m é o número de variáveis ou características dos elementos, x_{ij} corresponde ao valor da variável j para um elemento a testar i , x_j ao valor da variável j para o elemento observado e σ (sigma) é um parâmetro de dispersão da distribuição.

O único parâmetro ajustável da Equação 2.12 é o σ (sigma). Sua importância no desempenho da PNN é tal que para valores de σ muito pequeno, há uma tendência em supervalorizar os pontos amostrais, o que implica em pequena capacidade de generalização. Já para um valor de σ muito grande, o estimador tende a receber muita influência dos pontos vizinhos, o que implica em perda de informações do caso observado, ou seja, do ponto amostral (FALCÃO, 2002).

O ajuste dos parâmetros σ pode ser realizado sob diferentes abordagens em uma PNN: (1) utilização de apenas um valor de sigma para todo o modelo, (2) sigmas diferentes para variáveis diferentes, (3) sigmas diferentes para classes diferentes e (4) sigmas diferentes para variáveis e classes diferentes. O objetivo do ajuste em qualquer das abordagens é encontrar os valores de σ (sigma) que minimizem o erro de classificação, no entanto, a quarta abordagem é a que considera de maneira mais rigorosa a importância da influência que o valor sigma representa no desempenho da PNN. Assim, teoricamente pode-se garantir que construções de PNN que adotam a quarta abordagem para ajuste dos parâmetros σ apresentam maior desempenho que

aquelas que adotam as demais abordagens, notadamente a primeira delas. Na prática, a diferença de desempenho pode ser muito pequena em razão da maneira como se apresentam os dados utilizados para as construções das PNNs. Por exemplo, se todas as variáveis forem normalizadas sob um mesmo intervalo pode não fazer sentido a existência de diferentes sigmas para diferentes variáveis. No mesmo sentido, se a variação final dos casos de diferentes classes não for evidente pode não fazer sentido a existência de diferentes sigmas para diferentes classes.

2.7.2 Arquitetura

As PNNs possuem quatro camadas (Figura 2.6). O número de unidades de processamento na camada de entrada é igual ao número de características (ou atributos) de um elemento da amostra. A camada intermediária, também chamada camada padrão, é responsável pela implementação das regras de Bayes na qual as f.d.p. são aproximadas pelos estimadores de Parzen. Essa camada contém tantas unidades de processamento quanto são os casos da amostra de treinamento, pois cada unidade representa um padrão definido a partir de um caso do treinamento. A terceira camada, chamada camada de soma possui a mesma quantidade de unidades de processamento que o número de categorias ou classes desejadas. Cada unidade de processamento da terceira camada soma os valores resultantes dos padrões da camada anterior relativos a cada classe, a qual representa a probabilidade do elemento pertencer àquela classe. A classe com maior probabilidade é a selecionada para efeitos de classificação na camada de saída.

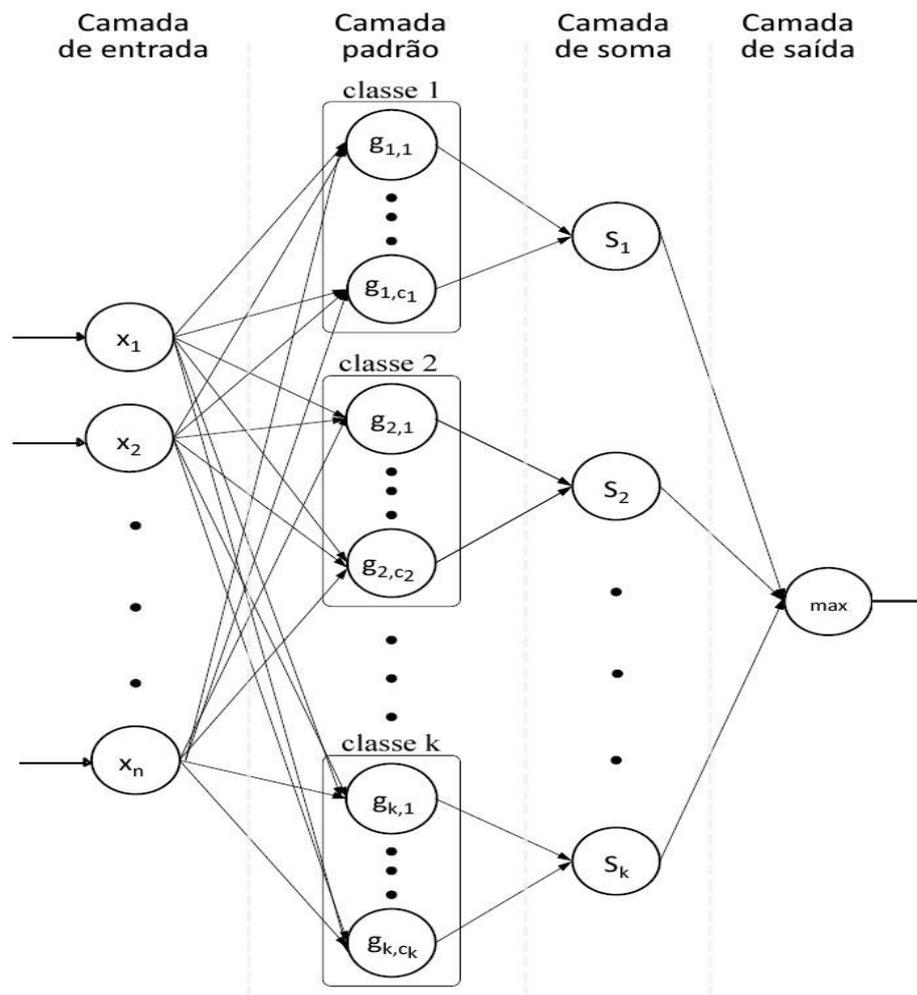


Figura 2.6 – Rede Neural Probabilística. Fonte: Adaptada de (GEORGIU *et al*, 2004).

2.7.3 Fluxo da Rede Neural Probabilística

Para realizar a classificação de um elemento da amostra, a PNN segue as seguintes etapas (FALCÃO, 2002):

- Um elemento da amostra é apresentado à camada de entrada de forma que o valor de cada variável do elemento é atribuído a um neurônio da camada de entrada;
- Calcula-se a distância de cada variável do elemento amostral em relação à variável correspondente em cada neurônio da camada padrão, através da aplicação da função núcleo;

- Cada neurônio calcula a probabilidade conjunta correspondente à distância que o elemento amostral de teste se encontra, atribuindo-lhe uma probabilidade de pertencer à sua própria classe;
- Essas probabilidades são somadas na camada de soma e escalonadas em relação ao número de neurônios de cada classe;
- O valor mais alto dos neurônios na camada de classificação corresponde à classe final.

Dessa forma, uma PNN faz a predição da classe de um novo elemento a partir da similaridade dele em relação a todos os elementos do conjunto de treinamento e a classificação final indica a classe a que pertencem os elementos de treinamento que são mais similares a esse novo elemento que está sendo testado, ponderando a quantidade de elementos de cada classe utilizada no treinamento. Por isso, diz-se que a PNN segue a abordagem de classificadores baseada em exemplos e, nesse caso, a adição de novos elementos ao conjunto de treinamento afetará gradualmente o resultado apresentado por modelos de classificação baseados nessa técnica.

3 Experimentos e Resultados

A verificação da possibilidade de que técnicas computacionais auxiliem a classificação de depósitos minerais pode ser feita pela aplicação de diferentes técnicas de aprendizado de máquina supervisionado sobre um conjunto de dados que disponha do conhecimento necessário aos propósitos pretendidos. A importância de tal classificação para o sistema de avaliação em três partes incentiva a utilização de informações que permitam identificar o modelo teor-tonelagem adequado à avaliação da região em estudo. As técnicas utilizadas, como já identificadas previamente, foram Árvore de Decisão, Rede Percéptron Multicamadas e Rede Neural Probabilística. Este capítulo apresenta a base de dados que serviu a pesquisa, bem como descreve os experimentos e resultados obtidos com o trabalho sobre as técnicas de aprendizado de máquina e os dados disponíveis.

3.1 Dados

A base de dados de depósitos minerais utilizada para desenvolvimento deste trabalho é composta por depósitos bem explorados com dados publicados na literatura (SINGER *et al*, 1993; SINGER *et al*, 2002; COX *et al*, 2003). São 1.861 depósitos que têm sua origem em diversas partes do mundo e estão classificados em tipos de acordo com os modelos descritos em (COX and SINGER, 1986; BLISS, 1992). Todos os depósitos de um determinado tipo compartilham um mesmo conjunto de regras. Por exemplo, a regra utilizada para determinar a área de um depósito em relação à possibilidade de combinar áreas mineralizadas próximas deve ser a mesma para todos os depósitos de um mesmo tipo. Assim, por exemplo, áreas de depósito do tipo *Porphyry copper* separadas por uma distância inferior a 2 quilômetros são consideradas como um depósito apenas e isso se aplica à delimitação de todos os depósitos do tipo *Porphyry copper* (SINGER *et al*, 2002).

O número de elementos na base de dados varia com o tipo de depósito, de acordo com os dados na Tabela 3.1. Os numerais que antecedem os nomes dos depósitos correspondem ao valor da variável que identifica o tipo de depósito na base de dados. Pode-se observar que há tipos com número de casos bem maior que outros tipos.

É, portanto, importante considerar essa distribuição na construção dos modelos de classificação e também na análise dos resultados.

Tabela 3.1 – Quantidade de depósito por tipo na base de dados.

Tipo de depósito	Quantidade
1 - <i>Porphyry Cu</i>	211
2 - <i>Porphyry Cu-Au</i>	113
3 - <i>Porphyry Cu-Mo</i>	52
4 - <i>Sediment-hosted Cu</i>	135
5 - <i>Cu skarn</i>	46
6 - <i>Zn-Pb skarn</i>	43
7 - <i>Polymetallic replacement</i>	78
8 - <i>Polymetallic vein</i>	50
9 - <i>Cyprus massive sulfide</i>	48
10 - <i>Kuroko massive sulfide</i>	424
11 - <i>Adularia-sericite Au-Ag</i>	131
12 - <i>Acid-sulfate Au-Ag</i>	28
13 - <i>Sediment-hosted Au</i>	27
14 - <i>Low-sulfide Au-quartz vein</i>	215
15 - <i>Homestake Au</i>	80
16 - <i>Sandstone-hosted Pb</i>	22
17 - <i>Sedimentary exhalative Zn-Pb</i>	128
18 - <i>Mississippi Valley Zn-Pb</i>	30
Total	1861

As características dos depósitos que compõem a base de dados são a tonelagem, a média do teor dos minerais cobre (Cu), molibdênio (Mo), prata (Ag), ouro (Au), zinco (Zn) e chumbo (Pb), e presença de 6 tipos de rocha: *carbonates, felsic to intermediate volcanic, granitoids, mafic volcanic, marine felsic to intermediate volcanic, marine mafic volcanic*. A tonelagem se refere à produção, reserva e recursos total de um depósito em seu mais baixo ponto de corte de teor.

Na publicação dos depósitos nem sempre estão acessíveis informações sobre os teores dos mesmos minerais, em especial quando se trata de tipos de depósitos diferentes. Da mesma forma, nem sempre está representado o tipo de rocha associado a eles. No entanto, ao se construir um modelo de classificação, é preciso que todos os casos compartilhem um mesmo conjunto de características. Além disso, muitas técnicas de aprendizado de máquina não estão aptas a lidar com a inexistência de valor para quaisquer das características. Assim, para garantir a existência de valores para as 13

características em cada depósito da base de dados, foram utilizadas as mesmas convenções adotadas em (SINGER, 2006): (1) teores não acessíveis foram tratados como zero; (2) depósitos sem indicação da ocorrência de rochas estão representados com zero em todos os seis tipos de rocha.

Devido ao fato de que as informações da maioria dos teores e tonelagem possuem uma distribuição log-normal (MARANHÃO, 1989) foi calculado o logaritmo desses valores antes de normalizá-los. Os valores dos teores foram somados a um para evitar o cálculo do logaritmo de zero. Em seguida, os dados foram normalizados através da aplicação da fórmula dada na Equação 3.1.

$$S_i = \frac{x_i - \mu}{\sigma} \quad \text{Eq. 3.1}$$

Os resultados S_i dessa normalização constam na base de dados e, juntamente com os valores 0 ou 1 sobre a presença ou ausência das rochas, formam o conjunto de 13 variáveis ou atributos que caracterizam os depósitos utilizados para construir e analisar o desempenho dos classificadores.

A base de dados aqui utilizada foi cedida pelo pesquisador Donald Singer e é similar à utilizada em (SINGER, 2006), porém com um número maior de elementos. Em relação àquela, o número de depósitos aumentou em 210 para o tipo *kuroko massive sulfide*, em 40 para *Adularia sericite Au-Ag* e em 1 elemento para *Low-sulfide Au-quartz vein*.

3.2 Metodologia dos Experimentos

Foi conduzida uma série de experimentos para realizar a análise dos classificadores e sua adequação ao domínio da prospecção mineral. O procedimento seguido em cada um deles é apresentado na Figura 3.1. Os dados da base foram utilizados para construir os modelos referentes a cada uma das técnicas e para efetuar testes nos modelos construídos, cujos resultados foram organizados em matrizes de confusão. A análise das matrizes de confusão oferece a justificativa das conclusões obtidas para cada experimento realizado.

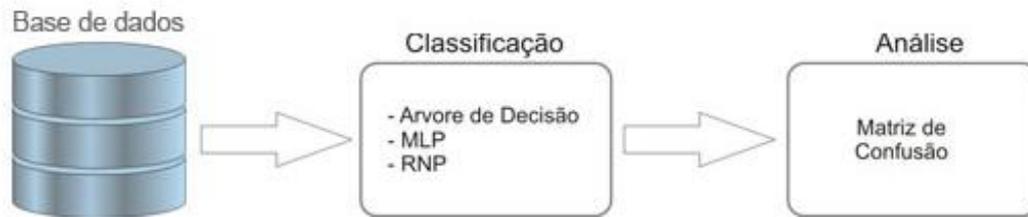


Figura 3.1 – Metodologia para comparação dos classificadores.

A realização da maioria dos experimentos seguiu o algoritmo da Listagem 3.1.

1. Divisão aleatória estratificada da amostra de dados do experimento em conjunto de treinamento e conjunto de teste na proporção de 50% dos elementos para cada conjunto;
2. Construção dos modelos de Árvore de Decisão, de Rede Percéptron Multicamadas e de Rede Neural Probabilística correspondentes ao conjunto de treinamento;
3. Aplicação do conjunto de teste aos modelos de Árvore de Decisão, de Rede Percéptron Multicamadas e de Rede Neural Probabilística gerados na etapa 2;
4. Armazenamento dos resultados encontrados na etapa 3 em uma matriz de confusão para cada um dos modelos de classificação;
5. Geração de uma matriz de confusão para cada técnica de aprendizado representando a porcentagem média de acertos/erros relativos de 100 execuções das etapas 1 a 4.

Listagem 3.1 – Algoritmo para realização dos experimentos.

Deve-se observar que o resultado final apresentado para análise é proveniente de 100 execuções das etapas que envolvem: a delimitação do conjunto de treinamento e conjunto de teste; a construção de modelos de classificação de cada técnica utilizada a partir do mesmo conjunto de treinamento; a aquisição de resultados proveniente da aplicação do conjunto de teste aos modelos gerados e a construção de uma matriz de confusão para organizar os resultados provenientes de cada modelo. Ainda, para maior

confiabilidade dos resultados, uma matriz de confusão final é gerada para cada técnica a partir dos resultados obtidos em 100 execuções das etapas acima.

Uma matriz de confusão representa um comparativo entre a classificação feita por especialistas e a classificação realizada pelo modelo gerado. A classificação dos especialistas é aquela que identifica o tipo de depósito na base de dados e a classificação do modelo corresponde ao tipo de depósito que o modelo consegue prever a partir das características que lhe são informadas. A disposição da matriz de confusão, com a classificação dos especialistas dispostos nas linhas e a classificação predita pelo modelo disposta nas colunas, permite verificar como o modelo de classificação distribui os depósitos de um determinado tipo identificado por especialistas dentre todos os tipos possíveis. A diagonal da matriz de confusão representa a concordância entre os especialistas e o modelo de classificação ou, dito de outra forma, a diagonal representa o acerto do modelo de classificação construído.

O desempenho geral de um classificador considerado neste trabalho é a razão entre o número de casos ou elementos corretamente classificados (independente do tipo) e o número de casos utilizados no conjunto de teste. Em razão de estarmos verificando a aplicabilidade das técnicas, utilizamos os termos desempenho da Árvore de Decisão, desempenho da Rede Percéptron Multicamadas e desempenho da Rede Neural Probabilística e estes representam a porcentagem de acerto médio de 100 modelos de classificação para cada técnica.

Dito isso, é importante perceber que uma análise mais minuciosa da aplicabilidade da técnica pode considerar sua eficácia em relação a um (ou mais) tipo de depósito em particular. Nesse caso o acerto considerado para um tipo de depósito é a razão entre o número de elementos classificados corretamente nesse tipo e o número de elementos desse tipo presentes no conjunto de teste. E, de forma similar, o desempenho da técnica em relação ao tipo particular é dado pelo acerto médio relativo a esse tipo.

As seções seguintes descrevem a maneira como foi feita a construção dos modelos de classificação de cada técnica utilizada.

3.2.1 Modelos de Árvore de Decisão

Os modelos de Árvore de Decisão foram construídos com base na implementação do algoritmo CART disponibilizada na biblioteca *Statistics Toolbox* do Matlab (MATHWORKS, 2010). Manteve-se o índice de Gini como critério para seleção de características e valores responsáveis pela divisão do nó. Isto foi feito em razão de este ser mais adequado quando há quantidades diferentes de elementos representando cada classe possível da amostra. Considerou-se o valor 5 como o mínimo de observações necessárias para criação de novo nó descendente. Os seis atributos relativos à presença de um tipo de rocha foram considerados atributos categóricos. Os sete demais foram considerados atributos contínuos ordenados.

Todas as árvores de classificação foram construídas com os critérios descritos e, a seguir, foram podadas no melhor nível estimado para elas. A estimação do melhor nível de poda considerou o erro mínimo calculado por um método de validação cruzada aplicado na árvore gerada sobre o conjunto de treinamento usado para sua construção. O método de validação utilizado, denominado *10-fold cross-validation*, é bastante utilizado para estimação do desempenho de classificadores (FUKUNAGA, 1990).

3.2.2 Modelos de Rede Percéptron Multicamadas

Os modelos de Rede Percéptron Multicamadas foram construídos com base na implementação disponibilizada pela biblioteca NETLAB (BISHOP *et al.*, 1996). Ela disponibiliza, entre outros, algoritmos para construção da MLP, para sua adaptação a um conjunto de dados (otimização) e para sua aplicação a um novo conjunto de dados. A construção de uma MLP adequada com essa biblioteca realiza duas etapas distintas: a preparação da estrutura da rede e sua adaptação ao conjunto de dados.

A construção da rede MLP necessita da identificação do valor dos seguintes parâmetros: número de entradas, número de unidades de processamento da camada oculta, número de unidades de processamento da camada de saída e a identificação da função de ativação que será utilizada na camada de saída. A estrutura resultante é uma rede *feedforward* de duas camadas com o número de neurônios informado para cada camada. Os pesos das conexões entre as entradas e os neurônios da camada oculta e das

conexões entre a camada oculta e a camada de saída são inicializados randomicamente. A função de ativação dos neurônios da camada oculta é sempre a tangente hiperbólica.

A otimização da MLP se refere à adaptação dos pesos entre as conexões da rede ao conjunto de treinamento através de otimizadores de propósito geral. Para essa etapa, é preciso determinar os seguintes parâmetros: o algoritmo de otimização que será utilizado e algumas opções que direcionam o seu comportamento, tais como o número máximo de iterações dos otimizadores, o valor de precisão absoluta requerida para encerrar o processo de otimização e o valor de precisão requerida para função objetivo. O número máximo de iterações dos otimizadores foi fixado em 500 em razão de ter-se verificado que o processo de otimização frequentemente finalizava antes de atingir esse número de iterações.

Em cada experimento, todos os modelos de classificação baseados em MLP utilizaram o mesmo conjunto de valores de parâmetros para sua construção e otimização. Os valores dos parâmetros foram previamente determinados de acordo com a amostra de dados do experimento. Cada parâmetro possui um conjunto de valores possíveis. A seleção da combinação de valores para o experimento foi realizada a partir do algoritmo da Listagem 3.2.

1. Divisão aleatória estratificada da amostra de dados do experimento em conjunto de treinamento e conjunto de teste na proporção de 50% dos elementos para cada conjunto;
2. Seleção de uma combinação de valores dos parâmetros;
3. Construção do modelo de Rede Percéptron Multicamadas correspondente ao conjunto de treinamento e à combinação de valores da etapa 2;
4. Aplicação do conjunto de teste ao modelo gerado na etapa 3;
5. Armazenamento da taxa de acerto resultante da etapa 3 junto à combinação de valores da etapa 2;
6. Execução das etapas 2 a 5 até que sejam realizadas todas as combinações de valores possíveis.

Listagem 3.2 – Algoritmo para determinação dos valores dos parâmetros adequados para construção de modelos de Rede Percéptron Multicamadas de um experimento.

Os parâmetros determinados para cada experimento foram: número de entradas, número de unidades de processamento da camada oculta, número de unidades de processamento da camada de saída, função de ativação da camada de saída, o algoritmo de otimização, o valor de precisão absoluta requerida para encerrar o processo de otimização e o valor de precisão requerida para função objetivo.

O número de entradas da MLP é sempre igual ao número de características que descrevem os casos. O número de unidades de processamentos da camada de saída é igual ao número de classes possíveis de resultarem do modelo de classificação. A determinação do algoritmo de otimização da MLP considerou as seguintes alternativas: gradiente conjugado, quase-newton e gradiente conjugado escalado. Uma apresentação desses algoritmos e sua utilização com NETLAB pode ser encontrada em (VICENTE, 2002).

Para determinar o número de unidades de processamento da camada oculta, o conjunto de valores possíveis compreendeu 10 valores igualmente distribuídos entre 5 e 50. Para a precisão absoluta e para a precisão da função objetivo, compreendeu 10 valores igualmente distribuídos de 0,01 a 0,1. Para a função de ativação dos neurônios da camada de saída, o conjunto de possíveis valores compreendeu as funções: linear, logística e softmax. A função de ativação linear apenas transmite o resultado do produto interno dos valores resultantes da camada oculta com os pesos das conexões entre a camada oculta e a camada de saída. A função logística aplica a Equação 2.10 ao produto interno descrito anteriormente. A função softmax é uma generalização da função logística para múltiplas variáveis e, de maneira similar àquela, permite que as saídas da rede sejam interpretadas como probabilidade *a posteriori*, com todos os valores de saída limitados entre zero e um.

A combinação de valores de todos os parâmetros que resultou na construção da MLP que apresentou melhor desempenho foi selecionada para o experimento.

3.2.3 Modelos de Rede Neural Probabilística

Os modelos de Rede Neural Probabilística foram construídos de acordo com a implementação presente na biblioteca *Neural Network Toolbox* do Matlab

(WASSERMAN, 1993). As PNNs construídas com essa implementação possuem apenas um único valor sigma para todo o modelo. O único parâmetro a ser estimado é o sigma, o qual foi determinado para cada experimento a partir do algoritmo da Listagem 3.3.

1. Divisão aleatória estratificada da amostra de dados do experimento em conjunto de treinamento e conjunto de teste na proporção de 50% dos elementos para cada conjunto;
2. Seleção de um valor para o parâmetro sigma;
3. Construção do modelo de Rede Neural Probabilística correspondente ao conjunto de treinamento e ao valor selecionado na etapa 2;
4. Aplicação do conjunto de teste ao modelo gerado na etapa 3;
5. Armazenamento da taxa de acerto resultante da etapa 3 junto ao valor selecionado na etapa 2;
6. Execução das etapas 2 a 5 até que sejam testados todos os valores possíveis.

Listagem 3.3 – Algoritmo para determinação do valor do parâmetro adequado para construção de modelos de Rede Neural Probabilística de um experimento.

O conjunto de valores testados para o parâmetro sigma compreendeu 100 valores igualmente distribuídos entre 0 e 1. O valor que resultou na construção da PNN que apresentou melhor desempenho foi selecionado para o experimento.

3.3 Experimento 1

O primeiro experimento realizado teve como objetivo verificar a aplicabilidade das técnicas de Árvores de Decisão, Rede Percéptron Multicamadas e Rede Neural Probabilística na classificação de depósitos minerais de acordo com os dados fornecidos na base de dados descrita na Seção 3.1 e com a distribuição dos elementos em conjunto de treinamento e conjunto de teste proposta na Tabela 3.2. Assim, os dados representam 18 classes ou tipos possíveis de depósitos minerais.

Tabela 3.2 - Distribuição dos depósitos da base de dados no conjunto de treinamento e no conjunto de teste.

Tipo de depósito	Treino	Teste
1 - <i>Porphyry Cu</i>	106	105
2 - <i>Porphyry Cu-Au</i>	57	56
3 - <i>Porphyry Cu-Mo</i>	26	26
4 - <i>Sediment-hosted Cu</i>	68	67
5 - <i>Cu skarn</i>	23	23
6 - <i>Zn-Pb skarn</i>	22	21
7 - <i>Polymetallic replacement</i>	39	39
8 - <i>Polymetallic vein</i>	25	25
9 - <i>Cyprus massive sulfide</i>	24	24
10 - <i>Kuroko massive sulfide</i>	212	212
11 - <i>Adularia-sericite Au-Ag</i>	66	65
12 - <i>Acid-sulfate Au-Ag</i>	14	14
13 - <i>Sediment-hosted Au</i>	14	13
14 - <i>Low-sulfide Au-quartz vein</i>	108	107
15 - <i>Homestake Au</i>	40	40
16 - <i>Sandstone-hosted Pb</i>	11	11
17 - <i>Sedimentary exhalative Zn-Pb</i>	64	64
18 - <i>Mississippi Valley Zn-Pb</i>	15	15
Total	934	927

Os modelos referentes à Árvore de Decisão foram gerados de acordo com a descrição na Seção 3.2.1 e, em cada execução do procedimento descrito na Listagem 3.1, utilizaram os elementos do conjunto de treinamento definido para aquela execução. A verificação do desempenho dos modelos de Árvore de Decisão foi feita através da análise da matriz de confusão final, a qual representa a média do desempenho dos 100 modelos gerados. A matriz de confusão final está apresentada na Tabela 3.3.

Tabela 3.3 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais em 18 tipos.

Tipo da base	Tipo de depósito predito pelo classificador																	
	Porphyry Cu	Porphyry Cu-Au	Porphyry Cu-Mo	Sediment-hosted Cu	Cu skarn	Zn-Pb skarn	Poly-metallic replacement	Poly-metallic vein	Cyprus massive sulfide	Kuroko massive sulfide	Adularia-sericite Au-Ag	Acid-sulfate Au-Ag	Sediment-hosted Au	Low-sulfide Au-quartz vein	Homestake Au	Sandstone-hosted Pb	Sed. exhalative Zn-Pb	Mississippi Valley Zn-Pb
Porphyry Cu	83	8	6	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Porphyry Cu-Au	17	77	0	2	1	0	0	0	1	0	1	0	0	0	0	0	0	0
Porphyry Cu-Mo	21	0	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sediment-hosted Cu	0	0	0	95	0	0	0	1	2	0	0	0	0	0	0	1	0	0
Cu skarn	16	6	0	0	66	1	2	4	0	1	0	0	1	0	0	0	0	0
Zn-Pb skarn	2	0	0	0	1	41	51	3	0	0	1	0	1	0	0	0	0	0
Polymetallic replacement	1	0	0	0	5	25	58	5	1	0	1	0	2	1	0	0	0	0
Polymetallic vein	0	0	0	3	1	2	11	61	1	1	7	1	1	5	1	4	1	0
Cyprus massive sulfide	7	1	0	0	7	2	1	1	67	8	1	0	0	1	0	1	3	0
Kuroko massive sulfide	0	0	0	0	0	0	0	0	3	96	0	0	0	0	0	0	1	0
Adularia-sericite Au-Ag	0	2	0	0	1	1	1	2	0	1	83	3	2	0	0	1	2	0
Acid-sulfate Au-Ag	1	10	0	2	2	0	0	0	0	1	40	42	0	0	0	0	0	0
Sediment-hosted Au	0	1	0	0	1	0	0	0	1	0	21	0	69	2	3	0	0	0
Low-sulfide Au-quartz vein	0	0	0	0	0	0	0	0	0	0	2	0	0	85	12	0	0	0
Homestake Au	0	0	0	0	0	0	0	0	0	0	2	0	1	42	55	0	0	0
Sandstone-hosted Pb	0	0	0	1	0	1	1	0	0	1	5	0	0	0	0	65	22	2
Sedimentary exhalative Zn-Pb	0	0	0	2	0	1	1	0	1	8	1	0	1	0	0	2	78	5
Mississippi Valley Zn-Pb	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	2	85	13

A análise da matriz de confusão apresentada na Tabela 3.3 possibilita identificar a existência de algumas tendências de classificação, isto é, o fato de que o classificador identifica como resultado para a classificação de alguns tipos de depósitos minerais um subconjunto de tipos de depósitos dentre todos os 18 possíveis.

À luz das dificuldades colocadas em (SINGER, 2006) para diferenciação de alguns tipos de depósitos entre si, a matriz de confusão permite verificar numericamente a veracidade dessas dificuldades para classificadores construídos com a técnica de Árvores de Decisão. De fato, os depósitos do tipo *Porphyry Cu*, quando mal classificados, são identificados como *Porphyry Cu-Au* (8%) ou como *Porphyry Cu-Mo* (6%). Em contrapartida, depósitos do tipo *Porphyry Cu-Au*, quando mal classificados, são identificados como *Porphyry Cu* (17%) e depósitos do tipo *Porphyry Cu-Mo*, quando mal identificados, são também identificados como *Porphyry Cu* (21%).

Outra dificuldade encontrada em (SINGER, 2006) ocorria na diferenciação dos tipos *Zn-Pb skarn* e *Polymetallic replacement*. A matriz de confusão da Tabela 3.3 mostra que modelos de Árvore de Decisão também apresentam essa dificuldade, com em média 51% dos depósitos *Zn-Pb skarn* erroneamente classificados como *Polymetallic replacement* e em média 25% dos depósitos do tipo *Polymetallic replacement* erroneamente classificados como *Zn-Pb skarn*.

As demais tendências de distribuição dos erros de classificação apresentadas em (Singer, 2006) também podem ser numericamente verificadas na Tabela 3.3. A maior parte dos erros na identificação do tipo *Kuroko massive sulfide* identifica os depósitos como *Cyprus massive sulfide* e vice-versa. Em média 40% dos depósitos do tipo *Acid-sulfate Au-Ag* erroneamente são identificados como do tipo *Adularia-sericite Au-Ag*. Também os tipos *Low-sulfide Au-quartz vein* e *Homestake Au* são difíceis de diferenciar, com em média 42% dos depósitos de tipo *Homestake Au* erroneamente classificados como *Low-sulfide Au-quartz vein* e com em média 12% dos depósitos do tipo *Low-sulfide Au-quartz vein* erroneamente classificados como *Homestake Au*. O tipo *Mississippi Valley Zn-Pb* permanece com dificuldades para ser identificado e a maioria dos depósitos são erroneamente classificados como do tipo *Sedimentary exhalative Zn-Pb* (85%) pelos modelos de Árvore de Decisão.

Os modelos de Rede Percéptron Multicamadas foram gerados a partir do conjunto de treinamento delimitado para cada execução do procedimento descrito na

Listagem 3.1. Os valores dos parâmetros de configuração necessários para construir todos os modelos desse experimento foram encontrados experimentalmente através da construção de MLPs com as diferentes configurações possíveis a partir do conjunto de valores possíveis, anteriormente delimitado, e seleção da configuração que gerou a MLP com a melhor taxa de acerto. O algoritmo utilizado para seleção dos valores adequados está descrito na Listagem 3.2. Os valores dos parâmetros utilizados estão apresentados na Tabela 3.4.

Tabela 3.4 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas para classificação em 18 tipos de depósitos.

Parâmetro	Valor
Número de entradas	13
Número de saídas	18
Número de neurônios da camada oculta	45
Função de ativação da camada de saída	Linear
Algoritmo de otimização	Gradiente conjugado
Precisão absoluta	0,02
Precisão da função objetivo	0,01

Em cada execução do procedimento descrito na Listagem 3.1, os elementos do conjunto de teste foram aplicados à MLP gerada e o resultado dessa aplicação foram organizados em uma matriz de confusão. A matriz de confusão final, média das 100 execuções, é apresentada na Tabela 3.5.

Tabela 3.5 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais em 18 tipos.

Tipo da base	Tipo de depósito predito pelo classificador																	
	Porphyry Cu	Porphyry Cu-Au	Porphyry Cu-Mo	Sediment-hosted Cu	Cu skarn	Zn-Pb skarn	Poly-metallic replacement	Poly-metallic vein	Cyprus massive sulfide	Kuroko massive sulfide	Adularia-sericite Au-Ag	Acid-sulfate Au-Ag	Sediment-hosted Au	Low-sulfide Au-quartz vein	Homestake Au	Sandstone-hosted Pb	Sed. exhalative Zn-Pb	Mississippi Valley Zn-Pb
Porphyry Cu	86	6	7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Porphyry Cu-Au	14	83	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
Porphyry Cu-Mo	61	0	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sediment-hosted Cu	0	0	0	96	1	0	0	1	2	0	0	0	0	0	0	0	0	0
Cu skarn	4	0	0	0	94	1	1	0	0	0	0	0	0	0	0	0	0	0
Zn-Pb skarn	0	0	0	0	3	51	45	0	0	0	0	0	0	0	0	0	0	0
Polymetallic replacement	1	0	0	0	5	15	78	1	0	0	0	0	0	0	0	0	0	0
Polymetallic vein	0	0	0	0	0	4	7	81	0	1	5	0	0	0	0	0	1	0
Cyprus massive sulfide	0	0	0	0	0	0	0	0	84	15	0	0	0	0	0	0	0	0
Kuroko massive sulfide	0	0	0	0	0	0	0	0	2	97	0	0	0	0	0	0	1	0
Adularia-sericite Au-Ag	0	2	0	0	0	0	0	0	0	0	92	2	1	0	0	0	1	0
Acid-sulfate Au-Ag	1	8	0	0	0	0	0	0	0	1	67	21	0	0	0	0	1	0
Sediment-hosted Au	0	1	0	0	0	0	0	0	0	0	0	0	96	0	2	0	0	0
Low-sulfide Au-quartz vein	0	0	0	0	0	0	0	0	0	0	0	0	0	91	8	0	0	0
Homestake Au	0	0	0	0	0	0	0	0	0	0	1	0	2	51	46	0	0	0
Sandstone-hosted Pb	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85	13	0
Sedimentary exhalative Zn-Pb	0	0	0	0	0	0	1	0	0	7	1	0	0	0	0	2	84	4
Mississippi Valley Zn-Pb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	80	19

A análise da matriz de confusão da Tabela 3.5 permite verificar que os modelos de Rede Percéptron Multicamadas também apresentam as tendências de distribuição de erros que ocorrem com os modelos de Árvore de Decisão. Assim, os tipos *Porphyry Cu*, *Porphyry Cu-Au* e *Porphyry Cu-Mo* são confundidos entre si pelos modelos de classificação. Em média 45% dos depósitos *Zn-Pb skarn* são erroneamente classificados como *Polymetallic replacement* e em média 15% dos depósitos do tipo *Polymetallic replacement* são erroneamente classificados como *Zn-Pb skarn*. A relação entre os tipos *Cyprus massive sulfide* e *Kuroko massive sulfide* se torna mais evidente com uma média de 15% dos depósitos do tipo *Cyprus massive sulfide* erroneamente classificados como *Kuroko massive sulfide*. Os modelos de MLP classificam em média 67% dos depósitos do tipo *Acid-sulfate Au-Ag* como *Adularia-sericite Au-Ag*. Os modelos de MLP identificam em média 51% dos depósitos do tipo *Homestake Au* como *Low-sulfide Au-quartz vein* e a maioria dos erros na classificação de depósitos do tipo *Low-sulfide Au-quartz vein* identificam o tipo *Homestake Au* como resultado. Ainda, a maior parte dos erros de classificação de depósitos do tipo *Mississippi Valley Zn-Pb* identificam o tipo *Sedimentary exhalative Zn-Pb* como resultado.

Os modelos de Rede Neural Probabilística foram gerados de acordo com a descrição da Seção 3.2.3 e utilizaram o conjunto de treinamento delimitado para cada execução do procedimento na Listagem 3.1. O valor do parâmetro sigma foi determinado em testes para o conjunto de dados desse experimento. O valor do parâmetro utilizado está apresentado na Tabela 3.6.

Tabela 3.6 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística para classificação em 18 tipos de depósitos.

Parâmetro	Valor
sigma	0,55

O resultado da aplicação do conjunto de teste ao modelo PNN de cada execução foi organizado em uma matriz de confusão. A matriz de confusão final, a qual representa a média do desempenho de 100 modelos de PNNs, é apresentada na Tabela 3.7.

Tabela 3.7 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais em 18 tipos.

Tipo da base	Tipo de depósito predito pelo classificador																	
	Porphyry Cu	Porphyry Cu-Au	Porphyry Cu-Mo	Sediment-hosted Cu	Cu skarn	Zn-Pb skarn	Poly-metallic replacement	Poly-metallic vein	Cyprus massive sulfide	Kuroko massive sulfide	Adularia-sericite Au-Ag	Acid-sulfate Au-Ag	Sediment-hosted Au	Low-sulfide Au-quartz vein	Homestake Au	Sandstone-hosted Pb	Sed. exhalative Zn-Pb	Mississippi Valley Zn-Pb
Porphyry Cu	83	11	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Porphyry Cu-Au	50	48	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Porphyry Cu-Mo	64	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sediment-hosted Cu	0	0	0	98	1	0	0	0	1	0	0	0	0	0	0	0	0	0
Cu skarn	4	0	0	4	86	1	2	0	1	2	0	0	0	0	0	0	0	0
Zn-Pb skarn	0	0	0	0	2	49	42	1	0	0	0	0	0	0	0	0	5	1
Polymetallic replacement	0	0	0	0	5	23	63	3	0	1	1	0	0	0	0	0	2	1
Polymetallic vein	0	0	0	0	0	5	13	75	0	2	4	0	0	0	1	0	0	0
Cyprus massive sulfide	0	0	0	1	0	0	0	0	79	20	0	0	0	0	0	0	0	0
Kuroko massive sulfide	0	0	0	0	0	0	0	0	2	97	0	0	0	0	0	0	0	0
Adularia-sericite Au-Ag	1	2	0	0	0	1	2	1	0	2	83	5	0	2	0	1	1	0
Acid-sulfate Au-Ag	1	7	0	3	2	0	0	0	0	1	60	27	0	0	0	0	0	0
Sediment-hosted Au	1	1	0	0	0	0	0	0	0	0	0	0	96	0	1	0	0	0
Low-sulfide Au-quartz vein	0	0	0	0	0	0	0	0	0	0	0	0	0	92	7	0	0	0
Homestake Au	0	0	0	0	0	0	0	0	0	0	1	0	0	55	44	0	0	0
Sandstone-hosted Pb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	87	11	2
Sedimentary exhalative Zn-Pb	0	0	0	1	0	1	2	1	0	12	0	0	0	0	0	3	71	9
Mississippi Valley Zn-Pb	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2	62	34

A matriz de confusão final para a técnica Rede Neural Probabilística também evidencia numericamente as tendências de classificação apresentadas em (SINGER, 2006). A maior parte dos depósitos classificados erroneamente dos tipos *Porphyry Cu-Au* e *Porphyry Cu-Mo* identificam o tipo *Porphyry Cu* como resultado. Em contrapartida, depósitos do tipo *Porphyry Cu*, quando mal classificados, identificam os tipos *Porphyry Cu-Au* e *Porphyry Cu-Mo*. Os classificadores identificam em média 42% dos depósitos do tipo *Polymetallic replacement* como *Zn-Pb skarn* e, em contrapartida, 23% dos depósitos de *Zn-Pb skarn* como *Polymetallic replacement*. Em média 20% dos depósitos de *Cyprus massive sulfide* são classificados como *Kuroko massive sulfide* e a maior parte dos erros para o tipo *Kuroko massive sulfide* resultam em *Cyprus massive sulfide*. A maior parte dos depósitos do tipo *Acid-sulfate Au-Ag* (60%) são classificados como *Adularia-sericite Au-Ag*. Modelos de PNN classificam em média 55% de depósitos *Homestake Au* como *Low-sulfide Au-quartz vein* e a maior parte dos erros para o tipo *Low-sulfide Au-quartz vein* resultam em *Homestake Au*. Para o tipo *Mississippi Valley Zn-Pb*, os modelos PNN identificam em média 62% de seus depósitos como *Sedimentary exhalative Zn-Pb*.

O Gráfico 3.1 apresenta as porcentagens de acerto para cada tipo de depósito resultante da aplicação dos classificadores. A Rede Percéptron Multicamadas foi a técnica que apresentou melhor resultado geral com uma média de acertos de 84% sobre os casos testados. A Rede Neural Probabilística apresentou um comportamento semelhante ao da MLP em relação ao acertos para cada tipo de depósito e cada modelo de PNN acertou em média 79% dos casos. Os modelos de Árvore de Decisão acertaram em média 79% dos casos.

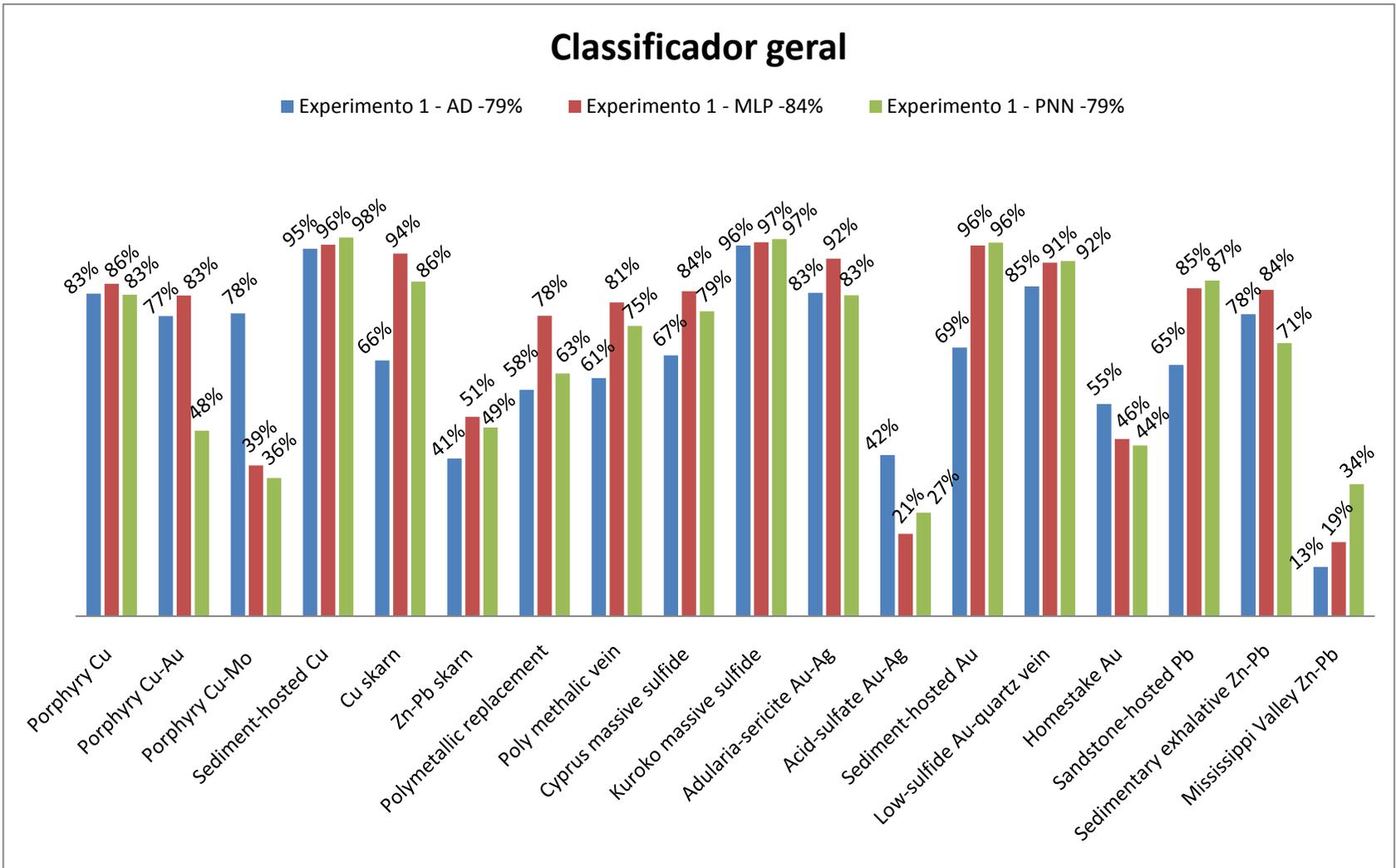


Gráfico 3.1 – Porcentagem de acertos dos classificadores considerando os 18 tipos de depósito.

A análise do Gráfico 3.1 permite identificar que a porcentagem de acerto para a maioria dos tipos de depósitos é semelhante para as técnicas Rede Percéptron Multicamadas e Rede Neural Probabilística. A essa afirmação merece ressalva a diferença das porcentagens de acerto para os tipos *Porphyry Cu-Au* (35%), *Polymetallic replacement* (15%), *Sedimentary exhalative Zn-Pb* (13%) e *Mississippi Valley Zn-Pb* (15%). Para os demais tipos essa diferença é inferior a 10%.

Em contrapartida, a atuação dos modelos de classificação gerados pela técnica de Árvore de Decisão mostra-se diferenciada para vários tipos de depósitos. A diferença entre o desempenho da Árvore de Decisão e o desempenho das demais técnicas é superior a 10% para 8 tipos de depósitos, com destaque para os tipos *Porphyry Cu-Mo* (38%), *Sediment-hosted Au* (33%), *Cu skarn* (20%) e *Sandstone-hosted Pb* (20%). Somente para os tipos *Porphyry Cu-Mo*, *Acid-sulfate Au-Ag* e *Homestake Au* o desempenho da Árvore de Decisão é melhor do que os das redes neurais aplicadas.

Embora existam diferenças na atuação de cada técnica, a análise das matrizes de confusão dessas três técnicas permite perceber a existência de um padrão de classificação semelhante ao descrito em (SINGER, 2006). Assim, pode-se citar as seguintes tendências de distribuição dos erros de classificação: (1) os tipos *Porphyry Cu*, *Porphyry Cu-Au* e *Porphyry Cu-Mo*, quando não são classificados corretamente tendem a ser confundido com um dentre esses três tipos; (2) a maioria dos erros de classificação dos tipos *Zn-Pb skarn* e *Polymetallic replacement* recaem em um dentre estes; (3) *Cyprus massive sulfide* e *Kuroko massive sulfide* são, da mesma forma, confundidos; (4) a maior parte dos erros para o tipo *Acid-sulfate Au-Ag* identificam o tipo *Adularia-sericite Au-Ag* como resultado; (5) os erros dos tipos *Low-sulfide Au-quartz vein* e *Homestake Au* identificam um dentre esses dois tipos como resultado; (6) a maior parte dos erros na classificação dos depósitos dos tipos *Mississippi Valley Zn-Pb* tendem a identificar o tipo *Sedimentary exhalative Zn-Pb* como resultado e o inverso ocorre com boa parte dos erros de classificação de depósitos do tipo *Sedimentary exhalative Zn-Pb*.

3.4 Experimento 2

Este segundo experimento teve como objetivo explorar o conhecimento proporcionado pela identificação, no Experimento 1 e em conformidade com a discussão em (SINGER, 2006), de um padrão na distribuição dos erros de classificação para alguns tipos de depósitos. Dessa forma, decidiu-se por verificar a atuação de classificadores gerados pelas técnicas de Árvores de Decisão, Rede Percéptron Multicamadas e Rede Neural Probabilística na classificação de depósitos minerais agrupados em tipos segundo a tendência apresentada ao final do Experimento 1. Foi feita uma modificação na variável que indica o tipo de depósito na base de dados para simular o agrupamento dos tipos. Dessa forma, formou-se os grupos apresentados na Tabela 3.8 através da substituição do valor da variável que identifica o tipo de depósito em todos os elementos de um grupo por um único valor. Assim, no Grupo 1, o valor dessa variável é 1 para todos os elementos dos tipos do grupo, no Grupo 6, o valor é 6, no Grupo 9, o valor é 9, no Grupo 11, o valor é 11, no Grupo 14, o valor é 14 e no Grupo 17, o valor é 17.

Tabela 3.8 – Grupo de tipos de depósitos considerados.

Grupo	Valor
Grupo 1	<i>Porphyry Cu</i> <i>Porphyry Cu-Au</i> <i>Porphyry Cu-Mo</i>
Grupo 6	<i>Zn-Pb skarn</i> <i>Polymetallic replacement</i>
Grupo 9	<i>Cyprus massive sulfide</i> <i>Kuroko massive sulfide</i>
Grupo 11	<i>Adularia-sericite Au-Ag</i> <i>Acid-sulfate Au-Ag</i>
Grupo 14	<i>Low-sulfide Au-quartz vein</i> <i>Homestake Au</i>
Grupo 17	<i>Sedimentary exhalative Zn-Pb</i> <i>Mississippi Valley Zn-Pb</i>

A distribuição dos dados em conjunto de treinamento e conjunto de teste proposta é apresentada na Tabela 3.9. Assim, os dados representam 11 classes, em que 5 dessas classes correspondem a tipos de depósitos minerais considerados nesta pesquisa

e 6 classes correspondem a um conjunto de tipos de depósitos minerais de acordo com a Tabela 3.9.

Tabela 3.9 – Distribuição dos depósitos agrupados no conjunto de treinamento e no conjunto de teste.

Tipo de depósito	Treino	Teste
1 - Grupo 1	188	188
4 - <i>Sediment-hosted Cu</i>	68	67
5 - <i>Cu skarn</i>	23	23
6 - Grupo 6	61	60
8 - <i>Polymetallic vein</i>	25	25
9 - Grupo 9	236	236
11 - Grupo 11	80	79
13 - <i>Sediment-hosted Au</i>	14	13
14 - Grupo 14	148	147
16 - <i>Sandstone-hosted Pb</i>	11	11
17 - Grupo 17	79	79
Total	933	928

Os modelos referentes à Árvore de Decisão foram gerados de maneira similar à sua geração no Experimento 1. A matriz de confusão apresentada na Tabela 3.10 representa as porcentagens médias de acertos e erros para cada tipo de depósito das 100 execuções.

Tabela 3.10 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais considerando a existência de grupos de depósitos.

Tipo da base	Tipo de depósito predito pelo classificador										
	Grupo 1	Sediment-hosted Cu	Cu skarn	Grupo 6	Polymetallic vein	Grupo 9	Grupo 11	Sediment-hosted Au	Grupo 14	Sandstone-hosted Pb	Grupo 17
Grupo 1	97	1	0	0	0	1	1	0	0	0	0
Sediment-hosted Cu	3	93	0	0	0	3	0	0	0	0	0
Cu skarn	12	0	77	3	3	2	2	0	1	0	0
Grupo 6	3	0	4	83	4	0	3	1	1	0	0
Polymetallic vein	0	3	1	13	62	2	7	1	5	4	2
Grupo 9	0	0	0	0	0	97	0	0	0	0	1
Grupo 11	5	2	1	2	2	1	84	1	0	0	2
Sediment-hosted Au	2	0	1	3	0	2	23	61	8	0	1
Grupo 14	0	0	0	0	0	0	2	0	97	1	0
Sandstone-hosted Pb	1	1	0	2	2	1	4	0	0	62	28
Grupo 17	0	1	0	2	0	7	1	1	0	1	85

Os modelos de Rede Percéptron Multicamadas foram gerados de maneira similar ao Experimento 1, porém foram utilizados os parâmetros de configuração presentes na Tabela 3.11. Esses parâmetros foram encontrados através de testes sobre os dados da base modificados nesse experimento para simular o agrupamento de tipos de depósitos. O procedimento para seleção dos valores de parâmetros adequados está descrito na Listagem 3.2.

Tabela 3.11 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando o agrupamento de tipos de depósitos.

Parâmetro	Valor
Número de entradas	13
Número de saídas	11
Número de neurônios da camada oculta	25
Função de ativação da camada de saída	Softmax
Algoritmo de otimização	Gradiente conjugado escalado
Precisão absoluta	0,02
Precisão da função objetivo	0,01

A matriz de confusão que representa o desempenho médio de 100 modelos de Rede Percéptron Multicamadas é apresentada na Tabela 3.12.

Tabela 3.12 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais considerando a existência de grupos de depósitos.

Tipo da base	Tipo de depósito predito pelo classificador										
	Grupo 1	Sediment-hosted Cu	Cu skarn	Grupo 6	Polymetallic vein	Grupo 9	Grupo 11	Sediment-hosted Au	Grupo 14	Sandstone-hosted Pb	Grupo 17
Grupo 1	98	0	1	0	0	0	1	0	0	0	0
Sediment-hosted Cu	0	97	1	0	0	2	0	0	0	0	0
Cu skarn	3	1	89	6	0	0	0	0	1	0	0
Grupo 6	0	0	3	92	2	0	0	0	0	0	1
Polymetallic vein	0	0	0	9	82	2	6	0	0	0	0
Grupo 9	0	0	0	0	0	99	0	0	0	0	1
Grupo 11	2	0	0	0	2	0	94	0	0	0	1
Sediment-hosted Au	0	0	0	0	0	0	1	96	3	0	0
Grupo 14	0	0	0	0	0	0	0	1	99	0	0
Sandstone-hosted Pb	0	0	0	0	0	0	0	0	0	87	12
Grupo 17	0	1	0	1	1	5	1	0	0	1	90

Os modelos de Rede Neural Probabilística foram gerados também em conformidade com a descrição da Seção 3.2.3. O valor do parâmetro sigma foi encontrado através de testes sobre a base de dados modificada desse experimento. O valor utilizado é apresentado na Tabela 3.13.

Tabela 3.13 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando o agrupamento tipos de depósitos.

Parâmetro	Valor
sigma	0,48

A matriz de confusão final, que resume o desempenho resultante de 100 modelos de PNN é apresentada na Tabela 3.14.

Tabela 3.14 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais considerando a existência de grupos de depósitos.

Tipo da base	Tipo de depósito predito pelo classificador										
	Grupo 1	Sediment-hosted Cu	Cu skarn	Grupo 6	Polymetallic vein	Grupo 9	Grupo 11	Sediment-hosted Au	Grupo 14	Sandstone-hosted Pb	Grupo 17
Grupo 1	99	0	0	0	0	0	0	0	0	0	0
Sediment-hosted Cu	0	98	1	0	0	1	0	0	0	0	0
Cu skarn	4	3	87	3	0	1	2	0	0	0	0
Grupo 6	0	0	4	88	2	1	1	0	0	0	4
Polymetallic vein	0	0	0	18	75	2	4	0	1	0	0
Grupo 9	0	0	0	0	0	99	0	0	0	0	0
Grupo 11	3	1	0	2	1	2	88	0	2	1	1
Sediment-hosted Au	2	0	0	0	0	0	0	96	2	0	0
Grupo 14	0	0	0	0	0	0	0	0	99	0	0
Sandstone-hosted Pb	0	0	0	0	0	0	0	0	0	86	14
Grupo 17	0	1	0	3	0	9	0	0	0	3	83

O Gráfico 3.2 facilita a comparação do desempenho das técnicas em relação à aplicação dos modelos de classificação aos dados considerando grupos de depósitos. A Rede Percéptron Multicamadas foi a técnica que apresentou um melhor resultado geral com o acerto 96% de casos testados quando considerados os 6 grupos de depósitos apresentados na Tabela 3.8. A Rede Neural Probabilística acertou em média 95% dos casos. Os modelos de Árvore de Decisão acertaram em média 91% dos casos.

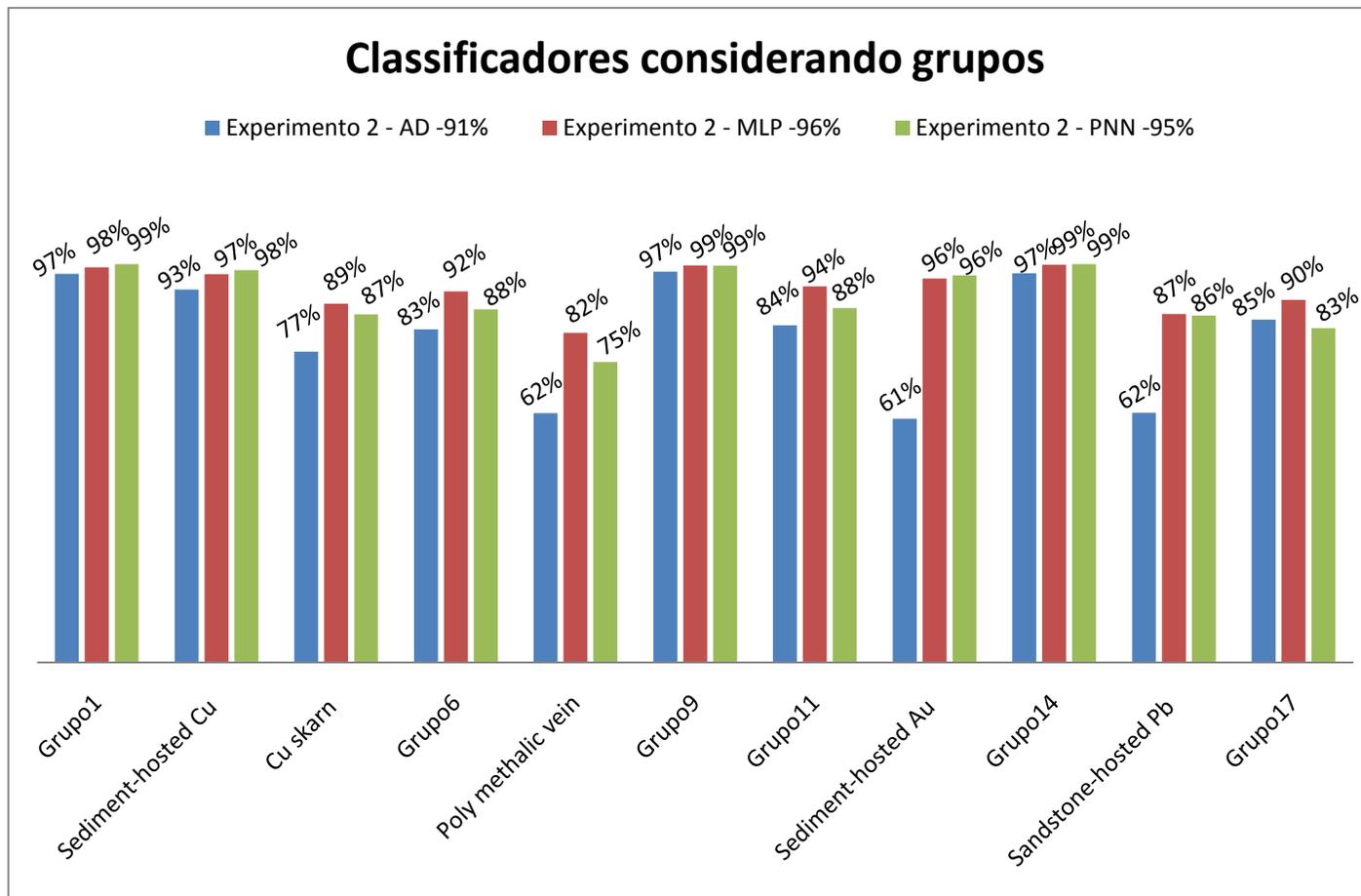


Gráfico 3.2 – Porcentagem de acertos dos classificadores considerando o agrupamento de depósitos.

É possível perceber que o acerto na classificação de depósitos de todos os grupos foi muito bom, independente da técnica utilizada. Para a Rede Percéptron Multicamada o acerto foi próximo a 100% para 3 grupos e nos demais foi superior a 90%. Ao comparar o Gráfico 3.2 com o Gráfico 3.1 pode-se verificar que para os tipos que continuaram a ser tratados individualmente não houve muita alteração nas porcentagens de acerto, sendo inferior a 6% para todos esses tipos.

O desempenho final de 96% para a MLP conseguida com o agrupamento dos dados versus 84% para a MLP que considerou os 18 tipos de depósitos individualmente permite acreditar que o principal desafio nessa classificação de depósitos minerais encontra-se na diferenciação dos depósitos dentro dos grupos formados. Esta verificação levantou alguns questionamentos: (1) seria possível que classificadores aplicados somente a um grupo conseguisse diferenciar melhor os tipos de depósitos desse grupo? (2) seria possível construir um modelo de classificação que inicialmente considerasse os grupos de depósitos e em seguida realizasse a diferenciação dos tipos de cada grupo? Os experimentos que se seguem buscam responder a esses questionamentos.

3.5 Experimento 3

Este terceiro experimento foi realizado com o objetivo de explorar a classificação de elementos pertencentes ao Grupo 1 de acordo com seus tipos específicos. Modelos de Árvore de Decisão, Rede Percéptron Multicamadas e Rede Neural Probabilística foram construídos a partir de um conjunto de treinamento formado somente por elementos de tipos do Grupo 1 e, em seguida, esses modelos foram aplicados a um conjunto de teste de modo a permitir a construção de matrizes de confusão que proporcionaram a análise das técnicas. A Tabela 3.15 identifica a quantidade de elementos utilizados nos conjuntos de treinamento e de teste.

Tabela 3.15 – Distribuição dos depósitos do Grupo 1 no conjunto de treinamento e no conjunto de teste.

Tipo de depósito	Treino	Teste
1 - <i>Porphyry Cu</i>	106	105
2 - <i>Porphyry Cu-Au</i>	57	56
3 - <i>Porphyry Cu-Mo</i>	26	26
Total	189	187

Dessa forma, esse experimento inclui 3 classes - os 3 tipos de depósitos do grupo *porphyry* – e verifica se alguma das técnicas utilizadas é capaz de diferenciar esses tipos com um grau de confiabilidade aceitável e, nesse caso, qual delas apresenta melhor desempenho.

Para a verificação da aplicabilidade de Árvore de Decisão a esta situação, os modelos foram gerados de maneira similar à sua geração no Experimento 1. A matriz de confusão apresentada na Tabela 3.16 representa as porcentagens médias de acertos e erros para cada tipo de depósito das 100 execuções.

Tabela 3.16 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 1.

	<u>Tipo de depósito predito pelo classificador</u>		
	<u>Porphyry Cu</u>	<u>Porphyry Cu-Au</u>	<u>Porphyry Cu-Mo</u>
<u>Tipo da base</u>			
Porphyry Cu	85	9	6
Porphyry Cu-Au	16	84	0
Porphyry Cu-Mo	22	0	78

A verificação da aplicabilidade de Rede Percéptron Multicamadas foi feita através da aplicação de modelos gerados de maneira similar aos experimentos anteriores com a configuração de seus parâmetros presente na Tabela 3.17.

Tabela 3.17 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 1.

Parâmetro	Valor
Número de entradas	13
Número de saídas	3
Número de neurônios da camada oculta	10
Função de ativação da camada de saída	Linear
Algoritmo de otimização	Gradiente conjugado
Precisão absoluta	0,02
Precisão da função objetivo	0,01

A matriz de confusão final que representa o desempenho médio resultante de 100 modelos MLPs é exibida na Tabela 3.18.

Tabela 3.18 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 1.

Tipo da base	Tipo de depósito predito pelo classificador		
	Porphyry Cu	Porphyry Cu-Au	Porphyry Cu-Mo
Porphyry Cu	86	6	8
Porphyry Cu-Au	14	86	0
Porphyry Cu-Mo	58	0	42

As PNNs foram construídas com o valor do parâmetro sigma apresentado na Tabela 3.19.

Tabela 3.19 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 1.

Parâmetro	Valor
sigma	0,79

A matriz de confusão que apresenta o desempenho médio dos modelos PNNs é apresentada na Tabela 3.20.

Tabela 3.20 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 1.

Tipo da base	Tipo de depósito predito pelo classificador		
	Porphyry Cu	Porphyry Cu-Au	Porphyry Cu-Mo
Porphyry Cu	91	5	4
Porphyry Cu-Au	65	35	0
Porphyry Cu-Mo	71	0	29

O Gráfico 3.3 evidencia as porcentagens de acerto para as três técnicas em relação à diferenciação de depósitos do Grupo 1. A Árvore de Decisão atuou melhor para esse caso apresentando um desempenho de 84%. A Rede Percéptron Multicamada acertou em média 80% dos casos. Os modelos de Rede Neural Probabilística acertaram em média 66% dos casos.

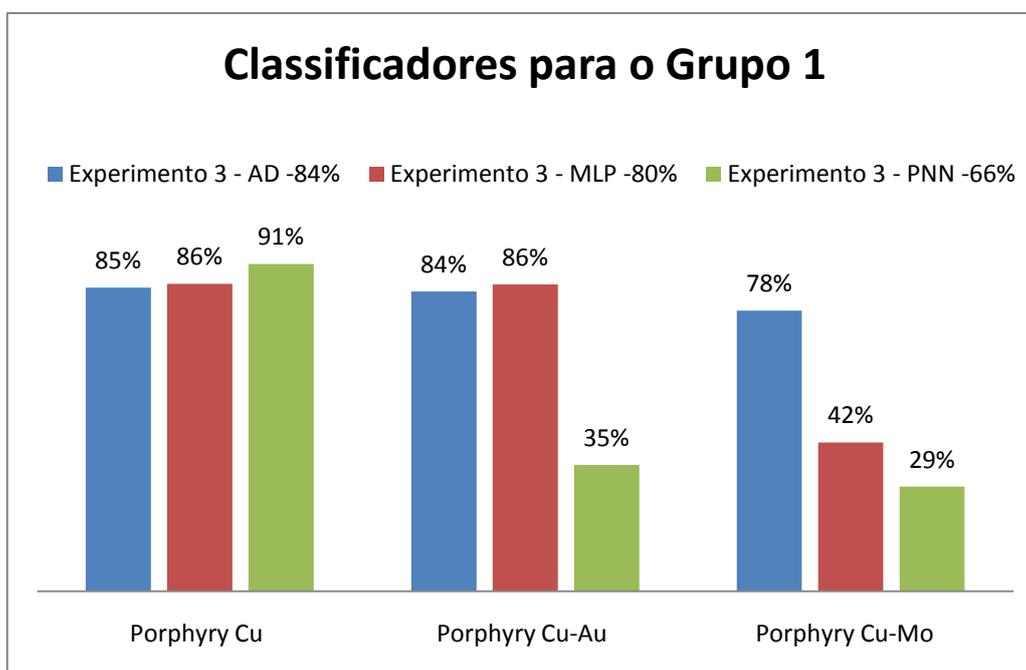


Gráfico 3.3 - Porcentagem de acertos dos classificadores para o Grupo 1.

Está evidente que o tipo *Porphyry Cu-Mo* é o mais difícil de ser diferenciado. Para esse tipo, árvores conseguem identificá-los em cerca de 78% dos casos, MLPs em 42% e PNNs em apenas 29% dos casos.

A análise das matrizes de confusão permite perceber que não existe confusão interna entre os tipos *Porphyry Cu-Au* e *Porphyry Cu-Mo*, mas apenas destes com o tipo *Porphyry Cu*. Desse modo, os erros do tipo *Porphyry Cu* se dividem entre os dois outros tipos, mas o erro daqueles sempre recai sobre o tipo *Porphyry Cu*.

A Tabela 3.21 mostra o acerto médio obtido para cada tipo do Grupo 1 para cada técnica e experimento realizado até o momento. A sua análise facilita a visualização da viabilidade de formação do Grupo 1.

Tabela 3.21 – Acerto médio obtido em cada aplicação para depósitos do Grupo 1.

Experimento	Técnica	<i>Porphyry Cu</i>	<i>Porphyry Cu-Au</i>	<i>Porphyry Cu-Mo</i>	Acerto médio
Experimento 1	AD	83%	77%	78%	81%
Experimento 1	MLP	86%	83%	39%	78%
Experimento 1	PNN	83%	48%	36%	66%
Experimento 3	AD	85%	84%	78%	84%
Experimento 3	MLP	86%	86%	42%	80%
Experimento 3	PNN	91%	35%	29%	66%

Pode-se perceber que a aplicação de Árvore de Decisão proporciona a melhor taxa de acerto para diferenciação entre os depósitos do Grupo 1 (84%) e que esta taxa é relativa ao experimento que considera a existência do Grupo 1. Os três tipos de depósito conseguem ser identificados neste caso, com destaque para o tipo *Porphyry Cu-Mo*, o qual somente consegue uma taxa de acerto aceitável com o uso de Árvore de Decisão.

3.6 Experimento 4

O quarto experimento foi realizado com o objetivo de explorar a classificação de elementos pertencentes ao Grupo 6, ou seja, para tratar a diferenciação entre os tipos *Zn-Pb skarn* e *Polymetallic replacement*. Modelos de Árvore de Decisão, Rede

Percéptron Multicamadas e Rede Neural Probabilística foram construídos a partir de um conjunto de treinamento formado somente por elementos de tipos do Grupo 6 e, em seguida, esses modelos foram aplicados a um conjunto de teste de modo a permitir a construção de matrizes de confusão que proporcionaram a análise das técnicas. A Tabela 3.22 identifica os dados utilizados neste experimento.

Tabela 3.22 – Distribuição dos depósitos do Grupo 6 no conjunto de treinamento e no conjunto de teste.

Tipo de depósito	Treino	Teste
6 - <i>Zn-Pb skarn</i>	22	21
7 - <i>Polymetallic replacement</i>	39	39
Total	61	60

Os modelos de Árvore de Decisão foram gerados de maneira similar aos experimentos anteriores, considerando apenas os dados deste experimento. A matriz de confusão apresentada na Tabela 3.23 representa as porcentagens médias de acertos e erros para cada tipo de depósito das 100 execuções.

Tabela 3.23 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 6.

	Tipo de depósito predito pelo classificador	
	Zn-Pb skarn	Polymetallic replacement
Tipo da base		
Zn-Pb skarn	8	92
Polymetallic replacement	6	94

A construção dos modelos de Rede Percéptron Multicamadas considerou a configuração colocada na Tabela 3.24. A identificação dos parâmetros adequados a esse experimento foi realizada da mesma forma que nos experimentos anteriores.

Tabela 3.24 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 6.

Parâmetro	Valor
Número de entradas	13
Número de saídas	2
Número de neurônios da camada oculta	30
Função de ativação da camada de saída	Linear
Algoritmo de otimização	Gradiente conjugado
Precisão absoluta	0,02
Precisão da função objetivo	0,01

A matriz de confusão final que representa o desempenho da técnica Rede Percéptron Multicamadas para esse experimento é exibida na Tabela 3.25.

Tabela 3.25 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 6.

	Tipo de depósito predito pelo classificador	
	Zn-Pb skarn	Polymetallic replacement
Tipo da base		
Zn-Pb skarn	51	49
Polymetallic replacement	30	70

As PNNs foram construídas com o valor do parâmetro sigma apresentado na Tabela 3.26, o qual foi selecionado pelo algoritmo descrito na Listagem 3.3

Tabela 3.26 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 6.

Parâmetro	Valor
sigma	0,83

A matriz de confusão que apresenta o desempenho médio dos modelos PNNs é apresentada na Tabela 3.27.

Tabela 3.27 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 6.

Tipo da base	Tipo de depósito predito pelo classificador	
	Zn-Pb skarn	Polymetallic replacement
Zn-Pb skarn	49	51
Polymetallic replacement	21	79

A atuação das técnicas na classificação de depósitos do Grupo 6 está evidenciada no Gráfico 3.4. É perceptível a dificuldade em distinguir os depósitos dos tipos *Zn-Pb skarn* e *Polymetallic replacement*. Nesta tarefa a Rede Neural Probabilística apresentou um desempenho um pouco melhor que as demais, com um acerto médio de 69% dos casos. A Rede Percéptron Multicamada acertou em média 63% dos casos e a Árvore de Decisão, em média 64% dos depósitos.

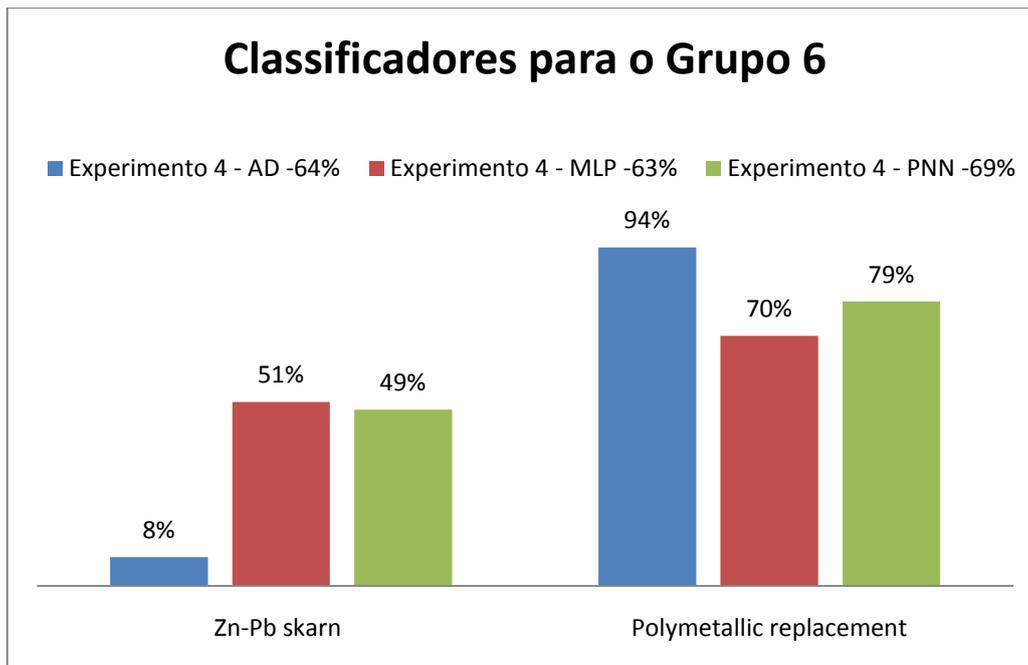


Gráfico 3.4 - Porcentagem de acertos dos classificadores para o Grupo 6.

A análise das matrizes de confusão permite perceber que as três técnicas tendem a classificar depósitos desse grupo como *Polymetallic replacement*. Existem aproximadamente 45% de depósitos do tipo *Polymetallic replacement* a mais que a quantidade de depósitos do tipo *Zn-Pb skarn* na base de dados. É comum que os classificadores associem depósitos que apresentam dúvidas na classificação ao tipo mais representativo na base de dados.

A Tabela 3.28 mostra o acerto médio obtido para cada tipo do Grupo 6 para cada técnica e experimento realizado até o momento. A sua análise facilita a visualização da viabilidade de formação do Grupo 6.

Tabela 3.28 – Acerto médio obtido em cada aplicação para depósitos do Grupo 6.

Experimento	Técnica	Zn-Pb skarn	Polymetallic replacement	Acerto médio
Experimento 1	AD	41%	58%	52%
Experimento 1	MLP	51%	78%	68%
Experimento 1	PNN	49%	63%	58%
Experimento 4	AD	8%	94%	64%
Experimento 4	MLP	51%	70%	63%
Experimento 4	PNN	49%	79%	69%

Pela Tabela 3.28 pode-se verificar que em todas as aplicações a taxa de acerto médio é inferior a 70%, o que indica que a utilização do Grupo 6 pode não ser vantajosa. Ainda assim, a maior taxa de acerto ocorre quando é considerada a existência do Grupo 6, com a aplicação de PNN (69%). Ao tratar os 18 tipos de depósitos individualmente, a MLP apresenta a maior taxa (68%).

3.7 Experimento 5

O quinto experimento foi realizado com o objetivo de verificar a classificação de depósitos dos tipos *Cyprus massive sulfide* e *Kuroko massive sulfide*. A quantidade de elementos utilizada neste experimento está na Tabela 3.29. São somente elementos do Grupo 9 que foram aleatoriamente divididos em conjunto de treinamento e de teste para realizarem o treinamento e a verificação das técnicas através de testes.

Tabela 3.29 – Distribuição dos depósitos do Grupo 9 no conjunto de treinamento e no conjunto de teste.

Tipo de depósito	Treino	Teste
9 - <i>Cyprus massive sulfide</i>	24	24
10 - <i>Kuroko massive sulfide</i>	212	212
Total	236	236

A matriz de confusão apresentada na Tabela 3.30 representa as porcentagens médias de acertos e erros obtidas pelos modelos de classificação de Árvore de Decisão gerados neste experimento.

Tabela 3.30 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 9.

Tipo da base	Tipo de depósito predito pelo classificador	
	Cyprus massive sulfide	Kuroko massive sulfide
Cyprus massive sulfide	88	12
Kuroko massive sulfide	4	96

A construção dos modelos de Rede Percéptron Multicamadas considerou a configuração colocada na Tabela 3.31. A identificação desses parâmetros foi realizada da mesma maneira dos experimentos anteriores.

Tabela 3.31 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 9.

Parâmetro	Valor
Número de entradas	13
Número de saídas	2
Número de neurônios da camada oculta	50
Função de ativação da camada de saída	Linear
Algoritmo de otimização	Gradiente conjugado
Precisão absoluta	0,02
Precisão da função objetivo	0,01

A Tabela 3.32 apresenta o desempenho da Rede Percéptron Multicamadas para esta aplicação.

Tabela 3.32 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 9.

Tipo da base	Tipo de depósito predito pelo classificador	
	Cyprus massive sulfide	Kuroko massive sulfide
Cyprus massive sulfide	84	16
Kuroko massive sulfide	2	98

Todas as Redes Neurais Probabilísticas construídas utilizaram o valor de sigma apresentado na Tabela 3.33, determinado da mesma forma que nos experimentos anteriores.

Tabela 3.33 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 9.

Parâmetro	Valor
sigma	0,4

A Tabela 3.34 mostra a matriz de confusão relativa à aplicação de Rede Neural Probabilística a este experimento.

Tabela 3.34 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 9.

Tipo da base	Tipo de depósito predito pelo classificador	
	Cyprus massive sulfide	Kuroko massive sulfide
Cyprus massive sulfide	82	18
Kuroko massive sulfide	2	98

O Gráfico 3.5 mostra as porcentagens de acerto média dos classificadores que representam as três técnicas utilizadas na diferenciação de depósitos do Grupo 9. A Rede Percéptron Multicamada atuou melhor para esse caso apresentando um

desempenho de 97%. A Rede Neural Probabilística apresentou taxa de acerto média 96%. Os modelos de Árvore de Decisão acertaram em média 95% dos casos.

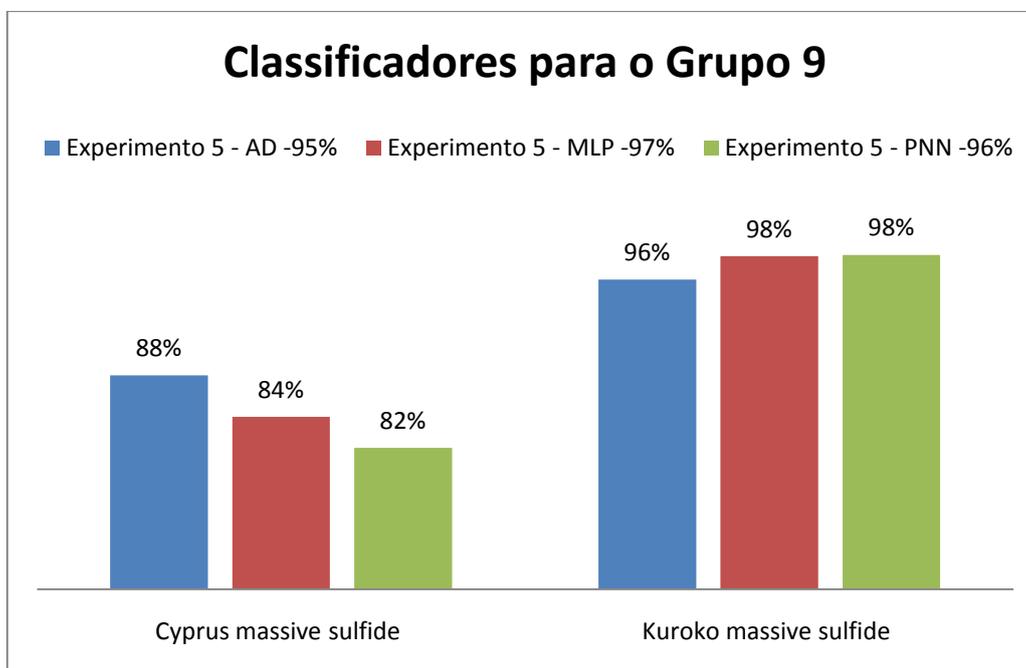


Gráfico 3.5 - Porcentagem de acertos dos classificadores para o Grupo 9.

A identificação do tipo *Kuroko massive sulfide* encontra acertos médios próximos a 100% para todas as técnicas. Os acertos para o tipo *Cyprus massive sulfide* são menores. Deve-se considerar a grande diferença entre a quantidade de elementos de cada tipo, pois o treinamento de classificadores do tipo *kuroko massive sulfide* é feito com uma quantidade de elementos quase dez vezes maior que os elementos usados do tipo *Cyprus massive sulfide*. Assim, os modelos de aprendizagem de máquina retêm um conhecimento maior sobre o tipo *Kuroko massive sulfide*, o que, normalmente, facilita a identificação de novos depósitos desse tipo.

A análise das matrizes de confusão corrobora as afirmações feitas anteriormente.

A Tabela 3.35 apresenta a taxa de acerto médio obtido para cada tipo do Grupo 9 relativa a cada técnica e experimento realizado até o momento. A sua análise facilita a visualização da viabilidade de formação do Grupo 9.

Tabela 3.35 – Acerto médio obtido em cada aplicação para depósitos do Grupo 9.

Experimento	Técnica	<i>Cyprus massive sulfide</i>	<i>Kuroko massive sulfide</i>	Acerto médio
Experimento 1	AD	67%	96%	93%
Experimento 1	MLP	84%	97%	95%
Experimento 1	PNN	79%	97%	95%
Experimento 5	AD	88%	96%	95%
Experimento 5	MLP	84%	98%	97%
Experimento 5	PNN	82%	98%	96%

Pode-se perceber as aplicações que consideraram a existência do Grupo 9 apresentaram melhores taxas de acerto que aquelas que tratou individualmente os 18 tipos de depósitos. As taxas de acerto foram bem próximas naqueles casos e a maior delas foi apresentada pela aplicação da MLP (97%).

3.8 Experimento 6

Este sexto experimento teve como objetivo verificar a melhor aplicação para diferenciar os tipos de depósitos do Grupo 11. Assim, como antes, são gerados modelos de Árvore de Decisão, Rede Percéptron Multicamadas e Rede Neural Probabilística. Os dados utilizados para treinamento e teste dos classificadores são apresentados na Tabela 3.36.

Tabela 3.36 – Distribuição dos depósitos do Grupo 11 no conjunto de treinamento e no conjunto de teste.

Tipo de depósito	Treino	Teste
11 - <i>Adularia-sericite Au-Ag</i>	66	65
12 - <i>Acid-sulfate Au-Ag</i>	14	14
Total	80	79

A matriz de confusão apresentada na Tabela 3.37 representa as porcentagens médias de acertos e erros obtidas pelos modelos de classificação de Árvore de Decisão gerados neste experimento.

Tabela 3.37 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 11.

Tipo da base	Tipo de depósito predito pelo classificador	
	Adularia-sericite Au-Ag	Acid-sulfate Au-Ag
Adularia-sericite Au-Ag	96	4
Acid-sulfate Au-Ag	12	88

A configuração apresentada na Tabela 3.38 foi utilizada para construir todos os modelos MLPs. A identificação desses parâmetros foi realizada da mesma maneira dos experimentos anteriores.

Tabela 3.38 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 11.

Parâmetro	Valor
Número de entradas	13
Número de saídas	2
Número de neurônios da camada oculta	20
Função de ativação da camada de saída	Linear
Algoritmo de otimização	Quase-Newton
Precisão absoluta	0,02
Precisão da função objetivo	0,01

A Tabela 3.39 apresenta o desempenho da Rede Percéptron Multicamadas para esta aplicação.

Tabela 3.39 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 11.

Tipo da base	Tipo de depósito predito pelo classificador	
	Adularia-sericite Au-Ag	Acid-sulfate Au-Ag
Adularia-sericite Au-Ag	94	6
Acid-sulfate Au-Ag	20	80

A Tabela 3.40 apresenta a configuração utilizada para construir todos os modelos de Rede Neural Probabilística. A identificação do parâmetro foi realizada da mesma maneira dos experimentos anteriores.

Tabela 3.40 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 11.

Parâmetro	Valor
sigma	0,7

A Tabela 3.41 apresenta a matriz de confusão com o desempenho da Rede Neural Probabilística para este experimento.

Tabela 3.41 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 11.

Tipo da base	Tipo de depósito predito pelo classificador	
	Adularia-sericite Au-Ag	Acid-sulfate Au-Ag
Adularia-sericite Au-Ag	96	4
Acid-sulfate Au-Ag	72	28

O Gráfico 3.6 mostra as porcentagens de acerto média dos classificadores que representam as três técnicas utilizadas na diferenciação de depósitos do Grupo 11. Nesta

aplicação a Árvore de Decisão apresenta a melhor taxa de acerto – 94%. A Rede Percéptron Multicamada apresentou um desempenho de 92% e a Rede Neural Probabilística de 84%.

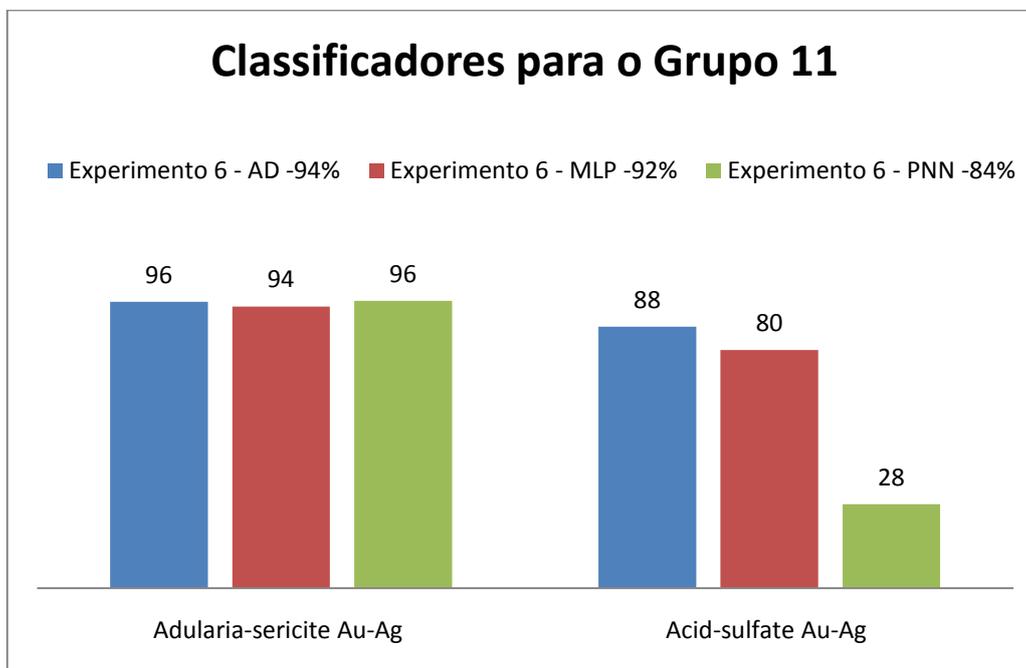


Gráfico 3.6 - Porcentagem de acertos dos classificadores para o Grupo 11.

A identificação do tipo *Acid-sulfate Au-Ag* encontra maior dificuldade e, para esse tipo, a Árvore de Decisão mostra-se mais eficiente. Deve-se ressaltar que o tipo *Adularia-sericite Au-Ag* possui quatro vezes mais representantes na base de dados do que o tipo *Acid-sulfate Au-Ag*.

A taxa de acerto médio obtido para cada tipo do Grupo 11 relativa a cada técnica e experimento realizado até o momento é apresentada na Tabela 3.42.

Tabela 3.42 – Acerto médio obtido em cada aplicação para depósitos do Grupo 11.

Experimento	Técnica	Adularia-sericite Au-Ag	Acid-sulfate Au-Ag	Acerto médio
Experimento 1	AD	83%	42%	76%
Experimento 1	MLP	92%	21%	80%
Experimento 1	PNN	83%	27%	73%
Experimento 6	AD	96%	88%	94%
Experimento 6	MLP	94%	80%	92%
Experimento 6	PNN	96%	28%	84%

As aplicações que consideraram a existência do Grupo 11 apresentaram melhores taxas de acerto que aquelas que tratou individualmente os 18 tipos de depósitos. A melhor taxa de acerto foi obtida pela aplicação de Árvore de Decisão considerando o Grupo 11 (94%).

3.9 Experimento 7

Este sétimo experimento teve como objetivo identificar a melhor técnica, dentre as utilizadas neste trabalho, para construir classificadores capazes de identificar os tipos específicos de depósitos do Grupo 14. Assim, como antes, são gerados modelos de Árvore de Decisão, Rede Percéptron Multicamadas e Rede Neural Probabilística. Os dados utilizados para treinamento e teste dos classificadores são o subconjunto de elementos da base de dados original formado apenas por depósitos dos tipos *Low-sulfide Au-quartz vein* e *Homestake Au*, tal como apresentado na Tabela 3.43.

Tabela 3.43 – Distribuição dos depósitos do Grupo 14 no conjunto de treinamento e no conjunto de teste.

Tipo de depósito	Treino	Teste
14 - <i>Low-sulfide Au-quartz vein</i>	108	107
15 - <i>Homestake Au</i>	40	40
Total	148	147

A Tabela 3.44 sintetiza a atuação da Árvore de Decisão a esse conjunto de dados. As árvores foram geradas da mesma forma que nos experimentos anteriores.

Tabela 3.44 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 14.

Tipo da base	Tipo de depósito predito pelo classificador	
	Low-sulfide Au-quartz vein	Homestake Au
Low-sulfide Au-quartz vein	87	13
Homestake Au	52	48

A configuração utilizada para MLPs seguiu a configuração que está na Tabela 3.45, a qual foi identificada através de testes tal como nos experimentos anteriores.

Tabela 3.45 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 14.

Parâmetro	Valor
Número de entradas	13
Número de saídas	2
Número de neurônios da camada oculta	40
Função de ativação da camada de saída	Linear
Algoritmo de otimização	Gradiente conjugado
Precisão absoluta	0,02
Precisão da função objetivo	0,01

O desempenho da Rede Percéptron Multicamadas para esta aplicação pode ser analisado pela matriz de confusão apresentada na Tabela 3.46.

Tabela 3.46 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 14.

Tipo da base	Tipo de depósito predito pelo classificador	
	Low-sulfide Au-quartz vein	Homestake Au
Low-sulfide Au-quartz vein	89	11
Homestake Au	46	55

A Tabela 3.47 apresenta a configuração utilizada para construir todos os modelos de Rede Neural Probabilística. A identificação do parâmetro foi realizada da mesma maneira dos experimentos anteriores.

Tabela 3.47 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 14.

Parâmetro	Valor
sigma	0,29

A matriz de confusão que apresenta o acerto médio de 100 modelos PNNs é apresentada na Tabela 3.48.

Tabela 3.48 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 14.

	Tipo de depósito predito pelo classificador	
	Low-sulfide Au-quartz vein	Homestake Au
Tipo da base		
Low-sulfide Au-quartz vein	88	12
Homestake Au	45	55

As porcentagens de acerto apresentadas no Gráfico 3.7 evidenciam o desempenho das técnicas na identificação de tipos de depósito do Grupo 14. De acordo com o que é apresentado, a Rede Percéptron Multicamadas tem desempenho um pouco melhor que as demais técnicas, embora o resultado final apresentado pelas três técnicas seja bastante semelhante. Elas identificam melhor depósitos *Low-sulfide Au-quartz vein*. O desempenho na classificação de depósitos do Grupo 14 foi de 80% para Rede Percéptron Multicamadas, 79% para Rede Neural Probabilística e 76% para Árvore de Decisão.

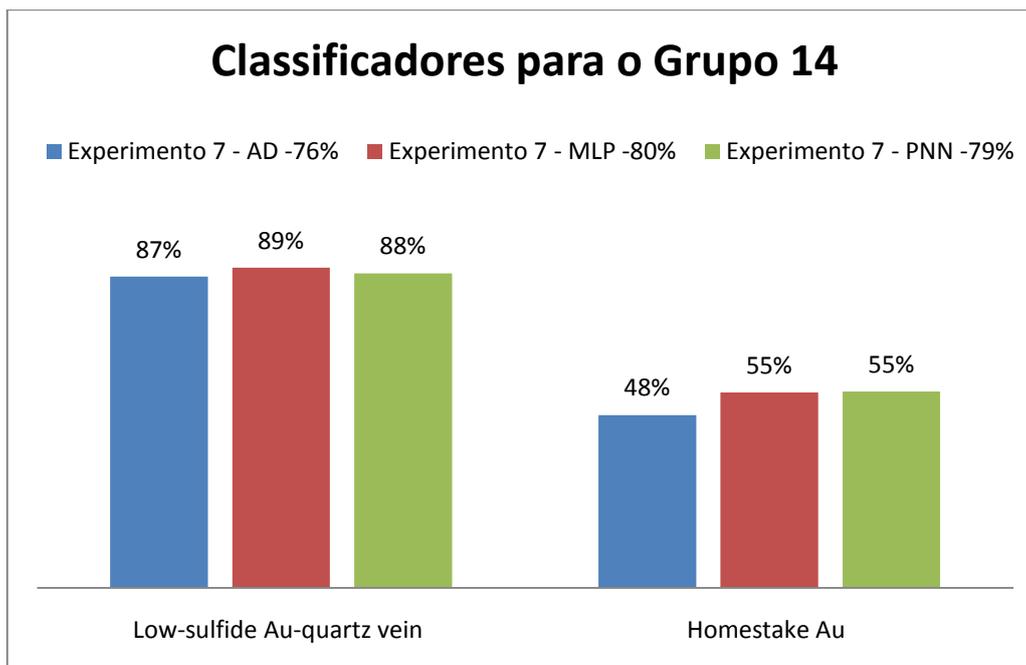


Gráfico 3.7 - Porcentagem de acertos dos classificadores para o Grupo 14.

Todas as técnicas encontram maior dificuldade para identificar o tipo *Homestake Au*, o qual apresenta-se em menor quantidade que o tipo *Low-sulfide Au-quartz vein* na base de dados.

A Tabela 3.49 resume o desempenho apresentado pelas técnicas utilizadas em cada experimento realizado até o momento.

Tabela 3.49 – Acerto médio obtido em cada aplicação para depósitos do Grupo 14.

Experimento	Técnica	<i>Low-sulfide Au-quartz vein</i>	<i>Homestake Au</i>	Acerto médio
Experimento 1	AD	85%	55%	77%
Experimento 1	MLP	91%	46%	79%
Experimento 1	PNN	92%	44%	79%
Experimento 7	AD	87%	48%	76%
Experimento 7	MLP	89%	55%	80%
Experimento 7	PNN	88%	55%	79%

As taxas de acerto das aplicações do experimento que considera o Grupo 14 são muito próximas das taxas conseguidas por aquelas que trataram os 18 tipos de

depósito de maneira individualizada. Ainda assim, a maior taxa (80%) foi obtida ao aplicar a técnica Rede Percéptron Multicamadas sobre o Grupo 14.

3.10 Experimento 8

O oitavo experimento se concentrou na busca da técnica, dentre as apresentadas neste trabalho, mais adequada para a classificação de depósitos que fazem parte do Grupo 17. Os dados neste experimento são, assim, aqueles da base de dados original que estão identificados com os tipos de depósitos *Sedimentary exhalative Zn-Pb* ou *Mississippi Valley Zn-Pb*. A delimitação dos dados é apresentada na Tabela 3.50.

Tabela 3.50 – Distribuição dos depósitos do Grupo 17 no conjunto de treinamento e no conjunto de teste.

Tipo de depósito	Treino	Teste
17 - <i>Sedimentary exhalative Zn-Pb</i>	64	64
18 - <i>Mississippi Valley Zn-Pb</i>	15	15
Total	79	79

A aplicação da técnica Árvore de Decisão foi realizada, como nos experimentos anteriores, com a construção de árvores e verificação do resultado da classificação de elementos do conjunto de teste. Os resultados obtidos estão resumidos na Tabela 3.51.

Tabela 3.51 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Árvore de Decisão para classificação de depósitos minerais do Grupo 17.

	Tipo de depósito predito pelo classificador	
	Sedimentary exhalative Zn-Pb	Mississippi Valley Zn-Pb
Tipo da base		
Sedimentary exhalative Zn-Pb	100	0
Mississippi Valley Zn-Pb	99	1

A construção de MLPs obedeceu a configuração apresentada na Tabela 3.52.

Tabela 3.52 – Configuração utilizada para construir classificadores da técnica Rede Percéptron Multicamadas considerando somente depósitos do Grupo 17.

Parâmetro	Valor
Número de entradas	13
Número de saídas	2
Número de neurônios da camada oculta	50
Função de ativação da camada de saída	Logística
Algoritmo de otimização	Gradiente conjugado escalado
Precisão absoluta	0,02
Precisão da função objetivo	0,01

A Tabela 3.53 organiza os resultados da aplicação desses modelos de classificação para possibilitar sua análise.

Tabela 3.53 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Percéptron Multicamadas para classificação de depósitos minerais do Grupo 17.

	Tipo de depósito predito pelo classificador	
	Sedimentary exhalative Zn-Pb	Mississippi Valley Zn-Pb
Tipo da base		
Sedimentary exhalative Zn-Pb	88	12
Mississippi Valley Zn-Pb	54	46

O valor do parâmetro sigma utilizado em todos os modelos PNN é apresentado na Tabela 3.54, o qual foi selecionado de acordo com o algoritmo apresentado na Listagem 3.3.

Tabela 3.54 – Configuração utilizada para construir classificadores da técnica Rede Neural Probabilística considerando somente depósitos do Grupo 17.

Parâmetro	Valor
sigma	0,5

A adequabilidade da Rede Neural Probabilística a essa classificação pode ser analisada a partir da matriz de confusão da Tabela 3.55, a qual apresenta a porcentagem média de acertos/erros provenientes do uso dos modelos de Rede Neural Probabilística construídos.

Tabela 3.55 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de Rede Neural Probabilística para classificação de depósitos minerais do Grupo 17.

Tipo da base	Tipo de depósito predito pelo classificador	
	Sedimentary exhalative Zn-Pb	Mississippi Valley Zn-Pb
Sedimentary exhalative Zn-Pb	88	12
Mississippi Valley Zn-Pb	61	39

As porcentagens de acerto apresentadas no Gráfico 3.8 evidenciam a dificuldade encontrada pela três técnicas para diferenciar depósitos do Grupo 17. A Árvore de Decisão resultou em 81% de acerto, a Rede Percéptron Multicamadas em 80% e a Rede Neural Probabilística em 79%. Os valores, inferiores a 50%, de acerto para o tipo *Mississippi Valley Zn-Pb* mostram a dificuldade de identificar esse tipo de depósito. Ao que parece, diante da decisão entre os tipos *Sedimentary exhalative Zn-Pb* e *Mississippi Valley Zn-Pb*, todos os classificadores tendem a escolher o primeiro tipo. Em especial a Árvore de Decisão erra quase a totalidade dos testes dos depósitos do tipo *Mississippi Valley Zn-Pb*.

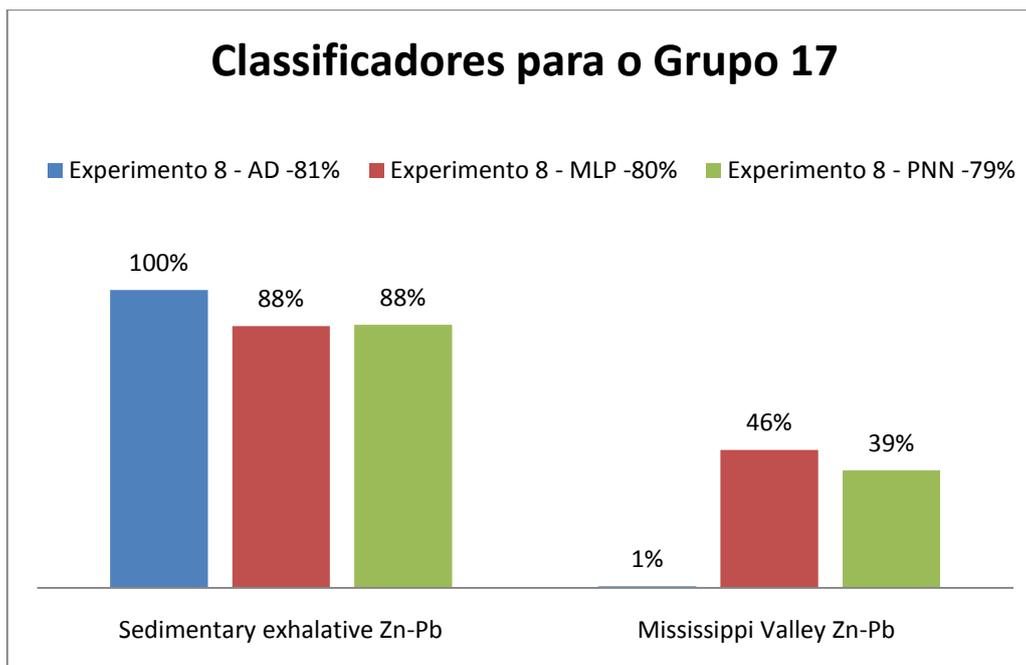


Gráfico 3.8 - Porcentagem de acertos dos classificadores para o Grupo 17.

A Tabela 3.56 resume o desempenho apresentado pelas técnicas utilizadas em cada experimento realizado até o momento.

Tabela 3.56 – Acerto médio obtido em cada aplicação para depósitos do Grupo 17.

Experimento	Técnica	<i>Sedimentary exhalative Zn-Pb</i>	<i>Mississippi Valley Zn-Pb</i>	Acerto médio
Experimento 1	AD	78	13	66%
Experimento 1	MLP	84	19	72%
Experimento 1	PNN	71	34	64%
Experimento 8	AD	100%	1%	81%
Experimento 8	MLP	88%	46%	80%
Experimento 8	PNN	88%	39%	79%

As taxas de acerto das aplicações do experimento que considera o Grupo 17 foram melhores que as conseguidas por aquelas que trataram os 18 tipos de depósito de maneira individualizada. A maior taxa (81%) foi obtida ao aplicar a técnica Árvore de Decisão, porém esta errou praticamente todos os testes para o tipo *Mississippi Valley Zn-Pb*. A Rede Percéptron Multicamadas teve uma taxa de acerto semelhante à Rede

Neural Probabilística e conseguiu identificar uma quantidade maior de depósitos do tipo *Mississippi Valley Zn-Pb*.

3.11 Experimento 9

Este nono experimento evidencia uma tentativa de construir um modelo de classificação mais adequado do que os gerados pelas três técnicas anteriormente apresentadas. O objetivo é melhorar a identificação dos 18 tipos de depósitos minerais a partir do conhecimento obtido da seqüência de experimentos anteriores. Isso significa que o agrupamento de depósitos deve ser considerado em um nível de classificação preliminar e, em seguida, será feita uma segunda classificação para os casos previamente identificados como parte de um dos grupos. A Figura 3.2 representa a estratégia adotada por este novo modelo de classificação.

A classificação preliminar é feita tal como no Experimento 2. Nesse momento, as saídas possíveis são um dentre 5 tipos de depósitos específicos ou um dentre 6 grupos de depósitos. As saídas que identificam grupos de depósitos são, a seguir, submetidas a uma nova etapa de classificação no intuito de se chegar a um tipo específico de depósito. Ao final, o modelo deve possibilitar a identificação de um dentre os 18 tipos de depósitos presentes na base de dados original.

Para realizar a classificação preliminar, utilizou-se a técnica que apresentou melhor resultado no Experimento 2, ou seja, Rede Percéptron Multicamadas. O classificador MLP gerado para essa etapa seguiu a mesma configuração descrita naquele experimento, de modo a manter a confiabilidade de uma média de 96% de acerto para esta classificação.

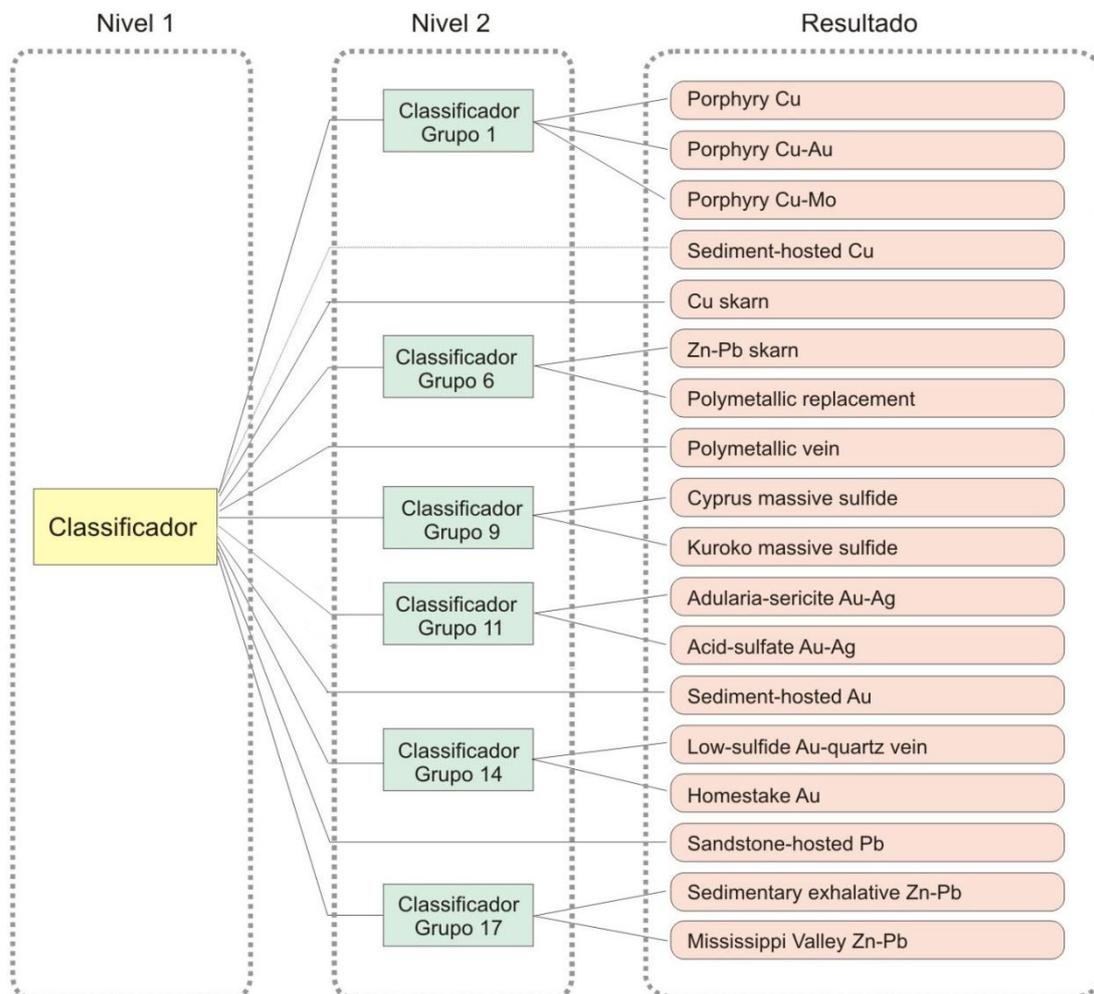


Figura 3.2 – Modelo de classificação formado pela combinação de classificadores organizados em 2 níveis de classificação.

Todos os casos que, nesta etapa preliminar, identificam um grupo de depósito como resultado devem ser direcionados a um novo classificador responsável pela diferenciação de depósitos em tipos específicos do grupo identificado. Assim, quando é identificado o Grupo 1, deve-se direcionar o caso para um classificador que o diferencie entre os tipos *Porphyry Cu*, *Porphyry Cu-Au* e *Porphyry Cu-Mo*. De maneira similar, casos que identificam os demais grupos devem ser direcionados para classificadores específicos daqueles grupos.

O classificador responsável pela diferenciação do Grupo 1 e do Grupo 11 são gerados pela técnica *Árvore de Decisão*, a qual apresenta melhor desempenho em relação às demais, como verificado no Experimento 3 e no Experimento 6. No mesmo sentido, para diferenciação de depósitos do Grupo 6, utiliza-se um modelo de Rede Neural Probabilística, de acordo com a exposição do Experimento 4. Ainda, casos previamente classificados no Grupo 9, Grupo 14 ou Grupo 17 utilizam, em uma segunda etapa de classificação, respectivamente, uma MLP construída tal como no Experimento 5, no Experimento 7 ou no Experimento 8.

A matriz de confusão resultante da classificação de depósitos utilizando o modelo de classificação em dois níveis proposto neste experimento é apresentada na Tabela 3.57. As porcentagens referem-se à média de erros/acertos derivados de 100 execuções dos procedimentos colocados no algoritmo apresentado na Listagem 3.4

1. Divisão aleatória estratificada da amostra de dados em conjunto de treinamento e conjunto de teste na proporção de 50% dos elementos para cada conjunto;
2. Construção do modelo de classificação em dois níveis utilizando o conjunto de treinamento;
3. Aplicação do conjunto de teste ao modelo gerado na etapa 2;
4. Armazenamento dos resultados encontrados na etapa 3 em uma matriz de confusão;
5. Geração de uma matriz de confusão final representando a porcentagem média de acertos/erros relativos de 100 execuções das etapas 1 a 4.

Listagem 3.4 – Algoritmo para realização do Experimento 9.

Tabela 3.57 – Matriz de confusão representando a porcentagem de acerto/erro resultante do teste realizado sobre modelos de classificação em dois níveis proposto no Experimento 9 para classificação de depósitos minerais em 18 tipos.

Tipo da base	Tipo de depósito predito pelo classificador																	
	Porphyry Cu	Porphyry Cu-Au	Porphyry Cu-Mo	Sediment-hosted Cu	Cu skarn	Zn-Pb skarn	Poly-metallic replacement	Poly-metallic vein	Cyprus massive sulfide	Kuroko massive sulfide	Adularia-sericite Au-Ag	Acid-sulfate Au-Ag	Sediment-hosted Au	Low-sulfide Au-quartz vein	Homestake Au	Sandstone-hosted Pb	Sed. exhalative Zn-Pb	Mississippi Valley Zn-Pb
Porphyry Cu	84	9	6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Porphyry Cu-Au	16	80	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0
Porphyry Cu-Mo	23	0	77	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sediment-hosted Cu	0	0	0	97	1	0	0	0	2	0	0	0	0	0	0	0	0	0
Cu skarn	3	0	0	1	89	3	3	0	0	0	0	0	0	0	0	0	0	0
Zn-Pb skarn	0	0	0	0	2	48	48	1	0	0	0	0	0	0	0	0	1	0
Polymetallic replacement	1	0	0	0	3	18	72	3	0	0	0	0	0	0	0	0	1	0
Polymetallic vein	0	0	0	0	0	2	7	82	0	2	6	0	0	0	0	0	0	0
Cyprus massive sulfide	1	0	0	0	0	0	0	0	83	16	0	0	0	0	0	0	0	0
Kuroko massive sulfide	0	0	0	0	0	0	0	0	2	97	0	0	0	0	0	0	1	0
Adularia-sericite Au-Ag	0	1	0	0	0	0	0	2	0	0	91	4	0	0	0	0	1	0
Acid-sulfate Au-Ag	0	7	0	2	0	0	0	0	0	0	11	79	0	0	0	0	0	0
Sediment-hosted Au	0	0	0	0	0	0	0	0	0	0	1	0	96	1	2	0	0	0
Low-sulfide Au-quartz vein	0	0	0	0	0	0	0	0	0	0	0	0	0	89	10	0	0	0
Homestake Au	0	0	0	0	0	0	0	0	0	0	1	0	1	45	53	0	0	0
Sandstone-hosted Pb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	87	11	1
Sedimentary exhalative Zn-Pb	0	0	0	1	0	0	1	1	0	6	0	0	0	0	0	2	76	12
Mississippi Valley Zn-Pb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55	44

3.12 Comparação dos Resultados dos Experimentos

A Tabela 3.58 apresenta o resultado geral para as quatro aplicações descritas neste trabalho que consideram modelos de classificação capazes de diferenciar os depósitos minerais em 18 tipos.

Tabela 3.58 – Taxa de acerto para os modelos de classificação que diferenciam os depósitos em 18 tipos para comparação dos resultados dos experimentos.

Aplicação	Taxa de Acerto
Experimento 1 – Árvore de Decisão	79%
Experimento 1 – Rede Percéptron Multicamada	84%
Experimento 1 – Rede Neural Probabilística	79%
Experimento 9 – Modelo de Dois Níveis	85%

O modelo de classificação que considera dois níveis de classificação apresentou o melhor resultado, o qual foi a taxa de acerto de 85% dos casos. A aplicação da técnica de Rede Percéptron Multicamadas obteve uma taxa de acerto bem próxima, 84%. Uma análise mais detalhada da modificação da taxa de acerto deve considerar sua variação em relação a cada tipo de depósito.

O Gráfico 3.9 permite comparar a aplicação dos diferentes modelos de classificação construídos para permitir a identificação de 18 diferentes tipos de depósitos minerais. As três primeiras legendas representam as técnicas de aprendizado aplicadas diretamente para a diferenciação dos 18 tipos de depósitos. A quarta legenda representa o modelo de classificação em dois níveis.

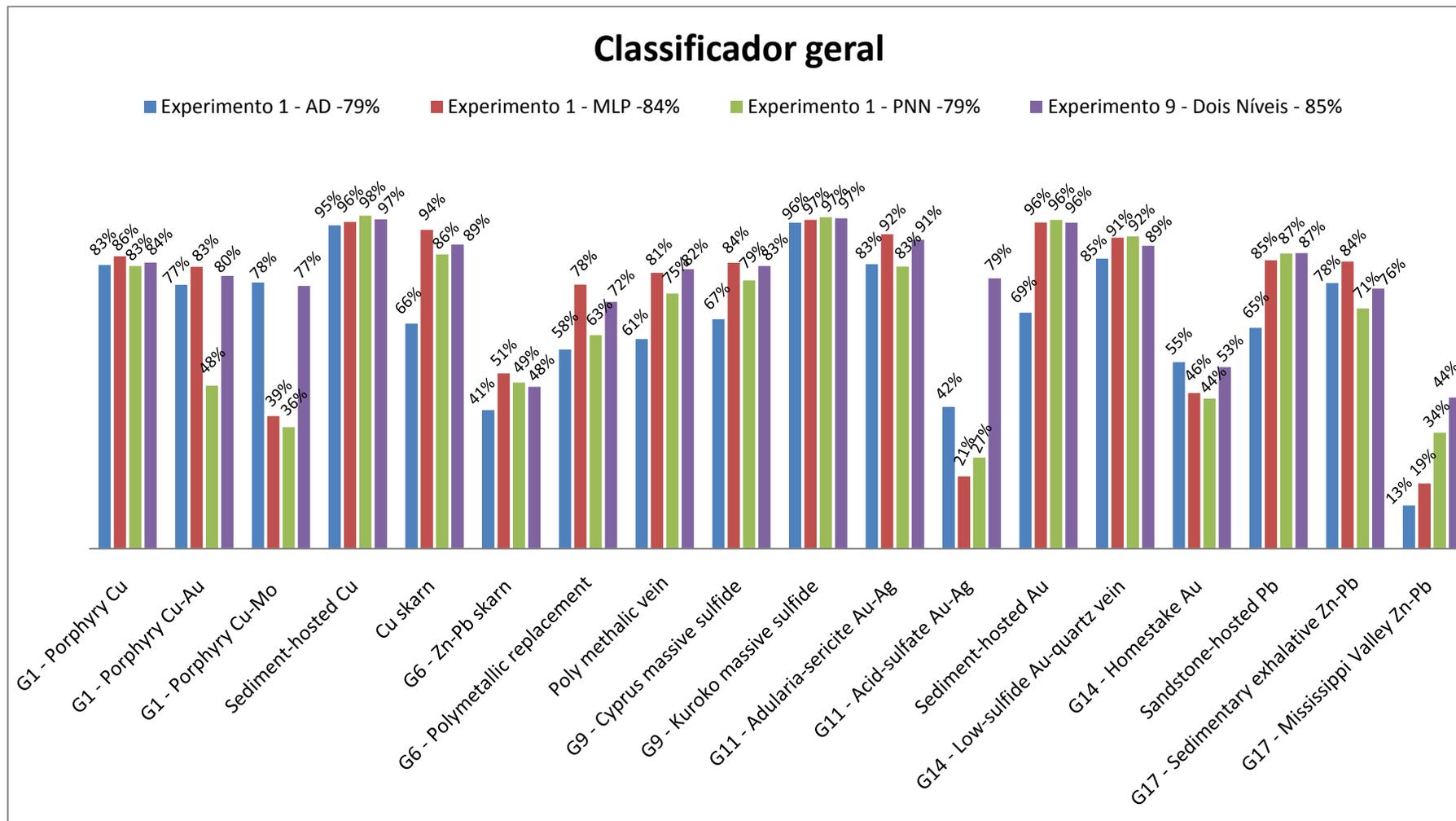


Gráfico 3.9 – Comparativo entre classificadores gerados com apenas uma técnica de aprendizado de máquina e sem considerar agrupamento de tipos e o modelo que considera dois níveis de classificação em razão do agrupamento de tipos de depósitos.

É possível verificar que a porcentagem de acerto relativa a alguns tipos de depósitos minerais apresenta melhor resultado no classificador construído com uma determinada técnica e para outros tipos o resultado pode ser melhor para o classificador que representa uma outra técnica. Isto sugere que a possibilidade de combinar diferentes técnicas pode melhorar o resultado final da classificação.

O modelo em dois níveis representa a combinação das técnicas estudadas neste trabalho a partir do reconhecimento da existência de grupos de depósitos. Os resultados de sua aplicação mostram que o modelo conseguiu apresentar desempenho melhor ou muito próximo ao melhor desempenho conseguido entre as três técnicas para a maioria os tipos de depósitos. A exceção fica para a diminuição na taxa de acerto dos tipos *Porphyry Cu-Au* (3%), *Cu skarn* (5%), *Zn-Pb skarn* (3%), *Polymetallic replacement* (6%), *Low sulfide Au-quartz vein* (3%), *Homestake Au* (2%) e *Sedimentary exhalative Zn-Pb* (8%).

A Tabela 3.59 apresenta a melhor aplicação para os tipos de depósitos que foram agrupados.

Tabela 3.59 – Melhor aplicação para os tipos de depósitos que foram agrupados para comparação dos resultados dos experimentos.

Grupo	Melhor Aplicação
Grupo 1	Experimento 9 – Modelo de Dois Níveis
Grupo 6	Experimento 1 – Rede Percéptron Multicamada
Grupo 9	Experimento 1 – Rede Percéptron Multicamada
Grupo 11	Experimento 9 – Modelo de Dois Níveis
Grupo 14	Experimento 9 – Modelo de Dois Níveis
Grupo 17	Experimento 9 – Modelo de Dois Níveis

A diminuição nas taxas de acerto do modelo em dois níveis em relação a aplicações que não consideraram grupos de depósitos para os tipos *Porphyry Cu*, *Porphyry Cu-Au* e *Porphyry Cu-Mo* é compensada pelo aumento da taxa de acerto quando considerado os três tipos de depósitos do Grupo 1. Raciocínio análogo pode ser aplicado às diminuições em relação aos tipos *Low-sulfide Au-quartz vein* e *Homestake Au*, os quais pertencem ao Grupo 14 no modelo de dois níveis.

O tipo *Sedimentary exhalative Zn-Pb* foi melhor identificado pela Árvore de Decisão, porém, essa mesma técnica prejudicou a identificação do outro tipo do Grupo 17, ou seja, o tipo *Mississippi Valley Zn-Pb*. Assim, para esse caso, deve-se considerar o ganho na classificação do Grupo 17 conseguida com o modelo em dois níveis, em detrimento à aparente perda da identificação do tipo específico. De forma análoga, o tipo *Adularia sericite Au-Ag* foi melhor identificado pela Rede Percéptron Multicamadas, porém essa técnica não conseguiu identificar bem o tipo *Acid-sulfate Au-Ag* e a consideração do Grupo 11 torna-se vantajosa.

Em relação ao desempenho na classificação dos tipos *Zn-Pb skarn* e *Polymetallic replacement*, a técnica de Rede Percéptron Multicamadas conseguiu obter melhor resultado em ambos os casos. Esses dois tipos foram reunidos como Grupo 6 na classificação em dois níveis. A confiabilidade da diferenciação desses tipos colocada no Experimento 4 (69% de acerto) e a atual verificação de que houve diminuição da taxa de acerto com a aplicação do modelo de dois níveis indica a necessidade de uma revisão da existência do Grupo 6, visto que a técnica de Rede Percéptron Multicamadas aplicada sem considerar os grupos conseguiu identificar melhor ambos os tipos. A existência do Grupo 9 também merece ser revista, pois também nesse caso a técnica de Rede Percéptron Multicamadas aplicada sem considerar os grupos conseguiu identificar um pouco melhor os tipos que pertencem a esse grupo.

A Tabela 3.60 apresenta a melhor aplicação para os tipos de depósitos que não foram agrupados.

Tabela 3.60 – Melhor aplicação para os tipos de depósitos que não foram agrupados para comparação dos resultados dos experimentos.

Tipo de Depósito	Melhor Aplicação
<i>Sediment-hosted Cu</i>	Experimento 1 – PNN: 1% melhor que Dois Níveis
<i>Cu skarn</i>	Experimento 1 – MLP: 5% melhor que Dois Níveis
<i>Polymetallic vein</i>	Experimento 9 – Modelo de Dois Níveis: 1% melhor que MLP
<i>Sediment-hosted Au</i>	Taxas iguais para: MLP, PNN e Dois Níveis
<i>Sandstone-hosted Pb</i>	Taxas iguais para: PNN e Dois Níveis

A diferença entre as taxas de acerto nas aplicações não foi muito grande, em média 5% no caso em que a diferença foi a maior. Pode-se perceber que apenas para um

desses tipos o modelo de dois níveis apresenta melhor resultado que as demais técnicas e essa diferença foi igual a 1%.

3.13 Comparação com Resultados Publicados

Uma pesquisa sobre aplicação de técnicas de aprendizado de máquina para a classificação de depósitos minerais em 18 tipos, tal como o utilizado neste trabalho, foi realizada em (SINGER, 2006) com aplicação de uma Rede Neural Probabilística. A Tabela 3.61 apresenta o resultado geral para a aplicação do modelo de dois níveis proposto neste trabalho e a aplicação apresentada em (SINGER, 2006).

Tabela 3.61 – Taxa de acerto para os modelos de classificação que diferenciam os depósitos em 18 tipos para comparação com resultados publicados.

Aplicação	Taxa de Acerto
Modelo de Dois Níveis	85%
Singer, 2006	82%

O desempenho geral do modelo de classificação conseguido neste trabalho mostra-se um pouco melhor que o conseguido naquele momento. A consideração do desempenho individual dos tipos de depósitos permite a identificação de diferenças e, portanto, de pontos de discussão sobre a melhor aplicação para esse caso de classificação.

O Gráfico 3.10 possibilita uma melhor visualização dos resultados obtidos neste trabalho comparado a resultados previamente publicados.

Classificador com níveis x Resultados publicados

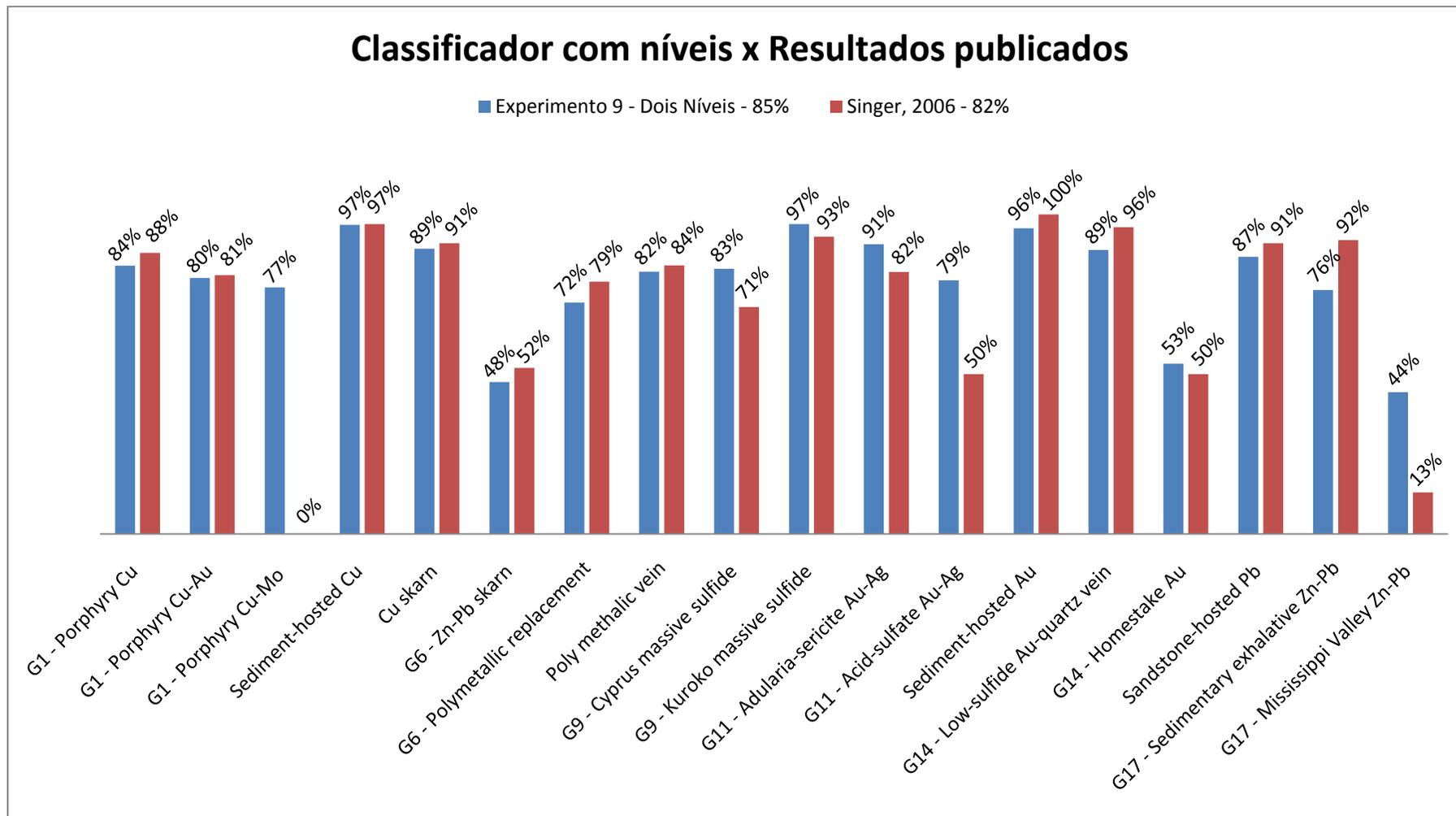


Gráfico 3.10 – Comparativo do resultado obtido com o modelo de classificação em dois níveis construído neste trabalho *versus* resultados publicados na literatura.

A Tabela 3.62 apresenta a melhor aplicação para os tipos de depósitos que foram agrupados.

Tabela 3.62 – Melhor aplicação para os tipos de depósitos que foram agrupados para comparação com resultados publicados.

Grupo	Melhor Aplicação
Grupo 1	Modelo de Dois Níveis
Grupo 6	Singer, 2006
Grupo 9	Modelo de Dois Níveis
Grupo 11	Modelo de Dois Níveis
Grupo 14	Modelo de Dois Níveis
Grupo 17	Modelo de Dois Níveis

Ao verificar a taxa de acerto de tipos que foram agrupados, claramente se percebe que o modelo de dois níveis apresentou melhores resultados para os grupos 1, 9 e 11. Em relação aos grupos 14 e 17, percebe-se que os tipos de depósito menos freqüentes na base de dados (*Homestake Au* e *Mississippi Valley Zn-Pb*) foram melhor identificados pelo modelo de dois níveis. Os tipos do Grupo 6 foram melhor identificados pela aplicação em (SINGER, 2006).

A Tabela 3.63 apresenta a melhor aplicação para os tipos de depósitos que não foram agrupados.

Tabela 3.63 – Melhor aplicação para os tipos de depósitos que não foram agrupados para comparação com resultados publicados.

Tipo de Depósito	Melhor Aplicação
<i>Sediment-hosted Cu</i>	Taxas iguais
<i>Cu skarn</i>	Singer, 2006: 2% melhor que Dois Níveis
<i>Polymetallic vein</i>	Singer, 2006: 2% melhor que Dois Níveis
<i>Sediment-hosted Au</i>	Singer, 2006: 4% melhor que Dois Níveis
<i>Sandstone-hosted Pb</i>	Singer, 2006: 4% melhor que Dois Níveis

Pode-se perceber que a maioria dos tipos que não fizeram parte de nenhum grupo é melhor identificada pela aplicação em (SINGER, 2006).

Os tipos que não fazem parte de nenhum grupo no modelo de classificação em dois níveis são identificados neste trabalho pela aplicação da Rede Percéptron Multicamadas no primeiro nível. Assim, uma alternativa que parece interessante é a substituição do modelo de classificação utilizado no primeiro nível por uma Rede Neural Probabilística configurada tal como em (SINGER, 2006). A aplicação de Rede Neural Probabilística usada naquele trabalho adota a abordagem de ajuste de sigma que considera valores de sigma diferentes para classes e variáveis diferentes, a qual é mais adequada para ajuste apropriado da rede neural aos dados da base de dados. A implementação de Rede Neural Probabilística utilizada neste trabalho permite a adoção de um único valor de sigma para todo o modelo e, portanto, adotou essa abordagem de ajuste.

A existência do Grupo 6 poderia ser reconsiderada se fosse utilizado PNN no primeiro nível, pois o desempenho conseguido por ambos os tipos mostrou-se melhor quando estes foram considerados isoladamente em (SINGER, 2006).

4 Conclusão

Nesta dissertação são apresentados alguns conceitos relacionados ao trabalho de prospecção mineral e a importância da classificação de depósitos minerais em tipos, de acordo com modelos de teor-tonelagem publicados. Apresentou-se diferentes tipos de modelos de depósitos minerais e sua utilidade na execução dos trabalhos de prospecção. A seguir, enfatizou-se a sistemática de avaliação quantitativa de depósitos em três partes e a necessidade de classificação de depósitos na aplicação dessa sistemática. Foi feita a contextualização da problemática de classificação de depósitos segundo a proposta de resolução de problemas de classificação com o uso de técnicas de aprendizado de máquina. A seguir foram apresentadas diferentes técnicas de aprendizado de máquina utilizadas para classificação de dados.

Para verificar a adequabilidade das técnicas apresentadas ao domínio delimitado nesse trabalho, foram realizados uma série de experimentos com a aplicação das técnicas de aprendizado supervisionado apresentadas sobre uma base de dados com informações dos teores, tonelagem e tipos de rochas associados a um conjunto de depósitos minerais previamente classificados por especialistas. A análise dos resultados dos experimentos permitiu a evolução do trabalho no sentido de construir um modelo de classificação considerando dois níveis de classificação, o que permitiu a melhoria dos resultados de classificação bem como a identificação de decisões mais críticas na classificação dos depósitos. Como resultado, é apresentada uma taxa de acerto de cerca de 85% para classificação dos depósitos em 18 tipos e a proveniência das principais ocorrências de erro são identificadas dentro dos grupos de tipos de depósito formados e em tipos de depósito menos representativos na base de dados, cujo número de elementos é inferior a 50.

No decorrer da realização dos experimentos, foram lançados dois desafios a este trabalho: o primeiro se refere à possibilidade de construir classificadores especializados na diferenciação de depósitos pertencentes a um grupo como os propostos neste trabalho e o segundo, à possibilidade de construir um único modelo para classificar os 18 tipos de depósitos de maneira a considerar as melhorias proporcionadas pelo agrupamento de depósitos.

O primeiro desafio foi assumido pelos experimentos 3 a 8. A melhor técnica para diferenciar tipos de depósitos de cada grupo foi identificada. Os resultados mostram que os grupos 1, 9 e 11 conseguem identificar o tipo específico dentro do grupo com taxas de acerto, respectivamente, de 84%, 97% e 94%. Os grupos 6, 14 e 17 mostraram-se mais difíceis de individualizar seus tipos. Os grupos 14 e 17 apresentaram taxas de acerto razoáveis, ambas com valor 80%. O grupo 6 mostrou-se ineficiente na diferenciação dentro do grupo com taxas de acerto 69%.

O segundo desafio é assumido pelo Experimento 9, o qual reúne o conhecimento adquirido pelos experimentos anteriores e propõe um modelo de classificação em dois níveis. O primeiro nível considera apenas 11 possíveis resultados à classificação, em que 6 desses resultados representam grupos de depósitos. Os testes que, no primeiro nível, resultaram em um grupo são, então, submetidos a um segundo nível de classificação, o qual identifica um tipo específico pertencente ao grupo. O modelo de classificação construído apresentou taxa de acerto de 85%, mas verificou-se que o acerto poderia melhorar se fosse utilizada uma implementação de Rede Neural Probabilística com valores de sigma para cada classe e variável no primeiro nível e fossem desconsiderados a existência dos grupos 6 e 9.

Como conclusão deste trabalho, podemos afirmar que não existem somente dois caminhos para a melhoria do processo de classificação de depósitos minerais como apresentado na literatura: aumento na quantidade de casos ou aumento na quantidade de características dos casos que fazem parte da base de dados.

Este trabalho apresenta um caminho promissor baseado na combinação de diversas técnicas de aprendizado de máquina que melhoram o resultado da classificação.

Fundamentadas nesta maneira alternativa de proporcionar melhoria na identificação de tipos de depósitos minerais, as sugestões para continuidade deste trabalho incluem a realização de novos experimentos com diferentes objetivos. Um objetivo seria a busca pela melhoria da classificação seguindo as sugestões propostas ao final dos comparativos dos experimentos ou mesmo através da utilização de outras técnicas de classificação em experimentos similares aos aqui colocados, pois pesquisas de aprendizado de máquina indicam que não é possível definir uma técnica de classificação ideal para todas as situações, mas apenas indicar aquela mais adequada para um domínio de aplicação particular.

Outro objetivo seria a verificação da aplicabilidade de métodos *ensemble*. Um *ensemble* designa um modelo de classificação cuja predição final resulta de predições realizadas por um conjunto de classificadores. Existem métodos definidos para geração do conjunto de classificadores que são a base para a construção do modelo final do *ensemble*. O trabalho seria, então, verificar se algum desses métodos gera modelos adequados para classificar os depósitos minerais.

Outra possibilidade inclui o estudo de novas estratégias de agrupamento de dados bem como de novas estratégias de combinação de classificadores. As estratégias de agrupamento de dados poderiam se basear, por exemplo, na análise de grupos formados pela aplicação de técnicas de aprendizado não supervisionado. As estratégias de combinação de classificadores poderiam incluir a investigação de abordagens e algoritmos de aprendizado de máquina propostos para a resolução de problemas hierárquicos, visto que se falou na formação de grupos de depósitos minerais e, portanto, na existência de uma hierarquia de classificação para identificar os tipos de depósitos pertencentes a um desses grupos formada por um classificador geral de primeiro nível e classificadores específicos no segundo nível.

Um caminho diferente seria a investigação de uma maneira alternativa para pré-processamento dos dados. Como colocado, as características disponibilizadas nas bases de dados publicadas na literatura podem ser diferentes para diferentes tipos de depósitos. Por exemplo, para o tipo *Sediment-hosted Au* apenas estão publicadas informações da tonelagem e dos teores de ouro e prata, mas para o tipo *Kuroko massive sulfide* são publicadas informações sobre a tonelagem e sobre os teores de cobre, prata, ouro, zinco e chumbo. Mesmo dentro dos tipos, algumas dessas informações podem não estar disponíveis. Nesse trabalho, foram adotadas as mesmas convenções utilizadas em (SINGER, 2006) para garantir o mesmo conjunto de informações para todos os depósitos da base de dados, ou seja, a atribuição do valor 0 (zero) para todas as características que não tinham valor determinado. Ao analisar os valores da base de dados, percebe-se que existem valores reais bem próximos a 0 (zero), de modo que existe a possibilidade de a convenção que foi adotada no pré-processamento estar influenciando o resultado da classificação. Uma alternativa seria a atribuição de valores que, ao final do pré-processamento, os diferenciassem de quaisquer outros valores existentes para aquelas características, o que poderia evitar sua influência sobre os resultados.

5 Referências bibliográficas

AMORIM, M. J. V.; BARONE, D.; MANSUR, A. U. (2008). Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Evasão Acadêmica. SBIE 2008, p. 666-674.

AMORIM, V. A.; ACAUAN, R. B.; STRIEDER, A. J.; JUCHEM, P. L. and HARTMANN, L. A. (2005). Controle Geológico-Estrutural dos Depósitos de Ametista no Distrito Mineiro de Ametista do Sul (RS). UNIVATES, Projeto de Referência 2784/03.

BARTON, P. B. (1986). Mineralogical Index *in* COX, D. P., and SINGER, D. A. Mineral Deposit Models. U. S. Geological Survey Bulletin 1693, p. 318-348.

BATISTA, A. J. (2006). Indicadores Ambientais na Pesquisa Mineral *in* Projeto BRA/01/039. Apoio à Reestruturação do Setor Energético. Programa das Nações Unidas para o Desenvolvimento, 24p.

BEHNIA, P. (2007). Application of Radial Basis Functional Link Networks to Exploration for Proterozoic Mineral Deposits in Central Iran. Natural Resources Research, v. 16, no. 2, p. 147-155.

BERK, R. A. (2006). An Introduction to Ensemble Methods for Data Analysis. Sociological Methods & Research, v. 34, p. 263-295.

BERRAR, D. P.; DOWNES, C. S.; DUBITZKY, W. (2003). Multiclass Cancer Classification Using Gene Expression Profiling and Probabilistic Neural Networks. Pacific Symposium on Biocomputing, v. 8, p. 5-16.

BISHOP, C. M. and NABNEY, I. T. (1996). NETLAB Online Reference Documentation. Disponível em <<https://ccrma.stanford.edu/~unjung/nethelp>>. Acessado em 30/03/2010.

BLISS, J. D. (1992). Developments in Mineral Deposit Modeling. U. S. Geological Survey Bulletin 2004, 168p.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A. and STONE, C. J. (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, California.

CAMARGO, L. S. and YONEYAMA, T. (2001). Specification or Training Sets and the Number of Hidden Neurons for Multilayer Perceptrons. Neural Computations, v. 13, p. 2673-2680.

COSTA, H. R. do N. (2006). Aplicação de Técnicas de Inteligência Artificial em Processos de Fabricação de Vidro. Tese de Doutorado, Escola Politécnica da Universidade de São Paulo, São Paulo, 248p.

COX, D. P. (1993). Estimation of Undiscovered Deposits in Quantitative Mineral Resource Assessments – Examples from Venezuela and Puerto Rico. Nonrenewable Resources, v. 2, no. 2, p. 82-91.

COX, D. P.; LINDSEY, D. A.; SINGER, D. A. and DIGGLES, M. F. (2003). Sediment-Hosted Copper Deposits of the World – Deposit Models and Database. U. S. Geological Survey Open-File Report 03-107.

COX, D. P. and SINGER, D. A. (1986). Mineral Deposit Models. U. S. Geological Survey Bulletin 1693, 379p.

DE'ATH, G. and FABRICIUS, K. E. (2000). Classification and Regression Trees: a Powerful Yet Simple Technique for Ecological Data Analysis. Ecology 81, v. 11, p. 3178-3192.

FALCÃO, A. (2002). Classificação de Ocupação Florestal em Fotografias Aéreas Orto-restituídas com Redes Neurais Probabilísticas. Grupo de Economia e Gestão de Recursos Naturais, Documento Técnico 03/02.

FLAKE, G. W. (1998). *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation*. MIT Press.

FUKUNAGA, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, 2ed.

GAMA, J. (2002). *Árvores de Decisão*. Palestra ministrada no Núcleo da Ciência de Computação, Universidade do Porto, Porto. Disponível em <<http://www.liaad.up.pt/~jgama>>. Acessado em 28/04/2010.

GEORGIU, V. L.; PAVLIDIS, N. G.; PARSOPOULOS, K. E.; ALEVIZOS, Ph. D.; VRAHATIS, M. N. (2004). Optimizing the Performance of Probabilistic Neural Networks in a Bioinformatics Task in Proceedings of the EUNITE 2004 Conference, p. 34-40.

HAMMARSTROM, J. M.; BRISKEY, J. A.; SCHULZ, K. J.; PETERS, S. G. and BAWIEC, W. J. (2006). Applications of a Global Mineral-Resource Assessment for Addressing Issues of Sustainable Mineral Resource Development *in* BARNHISEL, R. I. (2006). Proceedings. 7th International Conference on Acid Rock Drainage, American Society of Mining and Reclamation, Lexington, KY, p. 703-720.

HAYKIN, S. (2001). *Redes Neurais: Princípios e Práticas*. Trad. Paulo Martins Engel, 2 ed., Bookman, Porto Alegre.

KON, M. and PLASKOTA, L. (2000). Information Complexity of Neural Networks. *Neural Networks*, v. 13, p. 365-375.

KOTSIANTIS, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica Journal*, v. 31, no. 3, p. 249-268.

LEITE, E. P. e SOUZA FILHO, C. R. (2009). Artificial Neural Networks Applied to Mineral Potential Mapping for Copper-Gold Mineralizations in the Carajás Mineral Province, Brazil. *Geophysical Prospecting*, v. 57, p. 1049-1065.

MATHWORKS (2010). Statistics Toolbox User's Guide. The MarthWorks Inc. Disponível em <<http://www.mathworks.com/products/statistics/>>. Acessado em 30/03/2010.

MARANHÃO, R. J. L. (1989). Introdução à Pesquisa Mineral. 4ª ed. Imprensa Universitária. Fortaleza, BNB, ETENE. 796p.

MASTERS, T. (1995). Advanced algorithms for neural networks: A C++ Sourcebook. Wiley, New York, 425 p.

MCCAMMON, R. B. (1992). Numerical Mineral Deposit Models in BLISS, J. D. Developments in Deposit Modeling. U. S. Geological Survey Bulletin 2004, p. 6-12.

MILARÉ, C. R. (2003). Extração de Conhecimento de Redes Neurais Artificiais Utilizando Sistemas de Aprendizado Simbólico e Algoritmos Genéticos. Tese de Doutorado, Instituto de Ciências Matemáticas de São Carlos, São Carlos.

MONARD, M. C.; BATISTA, G. E. A. P. A.; PUGLIESI, J. B. (1997). Uma Introdução ao Aprendizado Simbólico de Máquina por Exemplos. ICMS – USP, São Carlos – SP.

MONTANA, D. (1992). A Weighted Probabilistic Neural Network. Advances in Neural Information Processing Systems 4, p.1110-1117.

NOEL, N. K. (2008). Intelligent Event Detection in Aircraft Engines. Tese de Mestrado, North Carolina State University, Raleigh.

PEREIRA, R. M. (2003). Fundamentos de Prospecção Mineral. Rio de Janeiro. Interciência.

PRÍNCIPE, J. C.; EULIANO, N. R.; LEFEBVRE, W.C. (2000). Neural and Adaptive Systems: Fundamentals through Simulations. Paperback, 672p.

SINGER, D. A. (1993). Basic Concepts in Three-Part Quantitative Assessments of Undiscovered Mineral Resources. Nonrenewable Resources, v. 2, no. 2, p. 69-81.

SINGER, D. A. (1996a). Introduction *in* SINGER D. A. An Analysis of Nevada's Metal-Bearing Mineral Resources. Nevada Bureau of Mines and Geology Open-File Report 96-2, Chapter 1, p.1.1-1.4.

SINGER, D. A. (1996b). Grade and Tonnage Models *in* SINGER D. A. An Analysis of Nevada's Metal-Bearing Mineral Resources. Nevada Bureau of Mines and Geology Open-File Report 96-2, Chapter 11, p. 11.1-11.18.

SINGER, D. A. (2006). Typing Mineral Deposits Using Their Associated Rocks, Grades and Tonnages Using a Probabilistic Neural Network. *Mathematical Geology*, v. 38, no. 4, p. 465-474.

SINGER, D. A. (2008a). Short Course Introduction to Quantitative Mineral Resource Assessments. U. S. Geological Survey, Open File Report 2007-1434.

SINGER, D. A. (2008b). Mineral Deposits Densities for Estimating Mineral Resources. *Mathematical Geosciences*, v. 40, no. 1, p. 33-46.

SINGER, D. A. (2010). Progress in Integrated Quantitative Mineral Resource Assessments. *Ore Geology Reviews*.

SINGER, D. A.; BERGER, V. I. and MORING, B. C. (2002). Porphyry Copper Deposits of the World – Database, Maps, and Preliminary Analysis. U. S. Geological Survey Open-File Report 02-268.

SINGER, D. A. and BERGER, V. I. (2007). Deposit Models and Their Application in Mineral Resource Assessments *in* BRISKEY, J. A. and SCHULZ, K. J. Proceedings for a Workshop on Deposit Modeling, Mineral Resource Assessment and Their Role in Sustainable Development. U. S. Geological Survey Circular 1294, p.71-78.

SINGER, D. A. and COX, D. P. (1988). Applications of Mineral Deposit Models to Resource Assessments. U. S. Geological Survey Bulletin, p. 55-57.

SINGER, D. A. and KOUDA, R. (1996). Application of a Feedforward Neural Network in the Search for Kuroko Deposits in the Hokuroku District. Japan. *Mathematical Geology*, v. 28, p. 1.17-1.23.

SINGER, D. A. and KOUDA, R. (1997a). Classification of Mineral Deposits into Types Using Mineralogy with a Probabilistic Neural Network. *Nonrenewable Resources*, v. 6, no. 1, p. 27-32.

SINGER, D. A. and KOUDA, R. (1999a). Examining Risk in Mineral Exploration. *Natural Resources Research*, v. 8, no. 2, p. 111-122.

SINGER, D. A. and KOUDA, R. (1999b). A Comparison of the Weights-of-Evidence Method and Probabilistic Neural Networks. *Natural Resources Research*, v. 8, no. 4.

SINGER, D. A. and KOUDA, R. (1999). A Comparison of the Weights-of-Evidence Method and Probabilistic Neural Networks. *Natural Resources Research*, v. 8, no. 4.

SINGER, D. A. and KOUDA, R. (2003). Typing Mineral Deposits Using Their Grades and Tonnages in an Artificial Neural Network. *Natural Resources Research*, v. 12, no. 3, p. 201-208.

SINGER, D. A.; MENZIE, W. D. and SUTPHIN, D. M.; MOSIER, D. L. and BLISS, J. D. (2001). Mineral Deposit Density – An Update. U. S. Geological Survey Professional Paper 1640.

SINGER, D. A.; MONSIE, D. L. and MENZIE, W. D. (1993). Digital Grade and Tonnage Data for 50 Types of Mineral Deposits – Macintosh Version. U. S. Geological Survey Open-File Report 93-280.

SOBRAL, A. P. B. (2003). Previsão de Carga-Horária: Uma Nova Abordagem por Árvore de Decisão. Tese de Doutorado, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 53p.

SPECHT, D. F. (1990). Probabilistic Neural Networks. *Neural Networks*, v. 3, p. 109-118.

VICENTE, T. (2002). *Redes Neurais para Inferência Estatística*. Universidade de São Paulo, São Paulo.

WASSERMAN, P. D. (1993). *Advanced Methods in Neural Computing*. New York, Van Nostrand Reinhold, p. 155-61.

YAN, G.; QIU, R.; LIAN C. *et al.* (2007). Quantitative Assessment of the Resource Potential of Porphyry Copper Systems in China. *Earth Science Frontiers*, v. 14, p. 27-41.

ZHAO, P. D. (2002). “Three-Component” Quantitative Resource Prediction and Assessments: Theory and Practice of Digital Mineral Prospecting. *Earth Science – Journal of China University of Geosciences*, v. 27, p. 482-489.

ZUO, R. Z.; CHENG, Q. and AGTERBERG, F. P. (2009). Application of a Hybrid Method Combining Multilevel Fuzzy Comprehensive Evaluation with Asymmetric Fuzzy Relation Analysis to Mapping Prospectivity. *Ore Geology Reviews*, v. 35, p. 101-108.