Modelo preditivo de património arqueológico

Natália Botica
Unidade de Arqueologia da Universidade do Minho, Braga, Portugal
nb@uaum.uminho.pt
Maribel Yasmina Santos
Departamento de Sistemas de Informação da Universidade do Minho, Guimarães, Portugal
maribel@dsi.uminho.pt
Francisco Sande Lemos
Unidade de Arqueologia da Universidade do Minho, Braga, Portugal
lemos@uaum.uminho.pt

Modelo preditivo de património arqueológico

Resumo

Os países europeus consideram que o legado arqueológico é um valor indissociável da sua identidade. Com o objectivo de evitar ou minimizar a destruição de bens arqueológicos estabeleceram instrumentos jurídicos e institucionais. No entanto, não basta preservar o Património conhecido, torna-se também necessário desenvolver planos preventivos para proteger sítios ocultos.

Os Sistemas de Informação são relevantes na concretização desta missão, apoiando o desenvolvimento de modelos que possam permitir a elaboração de cartas patrimoniais, onde esteja registado o património visível e oculto. O património oculto, ainda não descoberto, pode ser cartografado em função dos indicadores dos contextos e características dos sítios já referenciados e fornecidos pelos modelos preditivos, resultantes da Descoberta de Conhecimento em Bases de Dados.

O presente trabalho pretende ser um contributo para o desenvolvimento de cartas de risco, apresentando-se como caso de estudo para a elaboração de modelos preditivos de Património Arqueológico. Para tal, é utilizada uma Base de Dados que armazena um catálogo de sítios da região de Trás os Montes (TM), sobre a qual é aplicado o processo de Descoberta de Conhecimento em Bases de Dados. As técnicas de Data Mining utilizadas permitiram identificar modelos que sistematizam a localização de sítios Proto-históricos e Romanos.

Palavras chave: descoberta de conhecimento em bases de dados, data mining, construção de conhecimento arqueológico.

1. Introdução

A compreensão do passado passa pelo registo dos testemunhos que dele resultaram. A constituição de Bases de Dados do Património Arqueológico desde cedo despertaram o interesse dos arqueólogos, tendo estas crescido exponencialmente em volume e em número. O valor que lhes é atribuído também tem crescido, contribuindo para isso não só o seu intrínseco significado histórico e cultural, mas também o seu interesse para a indústria de turismo cultural, em franco desenvolvimento.

No entanto, não é de todo suficiente ter inventariado o património. É necessário retirar dessa inventariação o conhecimento que lhe é intrínseco. A Descoberta de Conhecimento em Bases de Dados tem encontrado aplicação nos mais variados domínios da ciência, constituindo a Arqueologia uma das áreas onde a sua aplicação poderá ser um valioso contributo.

Neste trabalho procura-se utilizar a Descoberta de Conhecimento em Bases de Dados Patrimoniais, para apoiar o desenvolvimento de algum conhecimento sobre o património inventariado e oculto, tendo em vista a sua preservação e mesmo posterior valorização. Dada a grande pressão urbanística, nomeadamente em meios urbanos, o património é descoberto não como resultado de sondagens e escavações planeadas mas, na maioria dos casos, em

consequência de empreendimentos públicos e privados cujos impactos patrimoniais são, muitas vezes, sub avaliados.

A avaliação dos impactos poderia ser mais precisa se, antes de qualquer intervenção no subsolo, fosse possível utilizar uma ferramenta de apoio indicativa do potencial património de uma determinada zona ou local.

Este trabalho pretende ser um contributo para este objectivo ao desenvolver um modelo que permita ao arqueólogo e às instituições da tutela prever a localização de arqueossítios.

O sistema de inferência de património arqueológico tem por base os princípios associados à Descoberta de Conhecimento em Bases de Dados (DCBD), na qual, através de um processo iterativo, se identificam relações entre os dados, com o intuito de construir modelos utilizáveis, por exemplo, em tarefas de previsão. O processo de DCBD inclui a utilização de técnicas de Data Mining (DM), cujos algoritmos permitem a identificação de padrões implícitos nos dados.

A incorporação do conhecimento arqueológico existente, ao longo do processo de validação do modelo, é constante e decisivo, sem o qual os modelos extraídos dos dados poderiam não constituir conhecimento arqueológico válido.

Este texto está estruturado em 5 secções. Na primeira faz-se um enquadramento do trabalho, definindo-se os objectivos e tarefas a desenvolver. Na segunda secção descreve-se sumariamente o processo de Descoberta de Conhecimento em Bases de Dados, reservando-se a terceira à aplicação desse processo aos sítios arqueológicos de Trás-os-Montes. Na quarta secção procede-se a uma análise e avaliação dos resultados obtidos, destinando-se a última secção para as conclusões e orientações em futuros trabalhos.

2. O Processo de Descoberta de Conhecimento em Base de Dados

As Bases de Dados e as Tecnologias da Informação que lhe estão associadas, são cada vez mais poderosas e sofisticadas, permitindo o armazenamento e utilização de grandes quantidades de dados. Graças a estas tecnologias, as Organizações têm visto o seu espólio de dados crescer exponencialmente. No entanto, apenas uma pequena parte destes dados é analisada e utilizada como instrumento de apoio à decisão, ou na formulação de hipóteses cognitivas. A restante é armazenada para garantir a sua posterior utilização, considerando que mais tarde poderá ser útil.

A convicção de que as bases de dados podem ser uma mais valia das Organizações, quando submetidas a um processo de análise e compreensão dos dados, fez emergir as ferramentas de descoberta do conhecimento.

O processo geral de descoberta de conhecimento, a partir de dados, designa-se por Descoberta de Conhecimento em Base de Dados (DCBD). Este processo recorre a técnicas de Data Mining (DM), aplicando aos dados algoritmos de extracção de padrões e incorporando conhecimento do

domínio de aplicação, através da interpretação adequada de resultados [Fayyad et al. 1996].
No processo de descoberta de conhecimento as técnicas de DM são utilizados para desenvolver actividades **descritivas** ou de **previsão** [Han e Kamber 2001].

Nas actividades **descritivas** o processo de DCBD é utilizado para extrair padrões que permitam a caracterização do comportamento dos dados e identificação de valores anómalos ou pouco usuais.

Numa actividade de **previsão** procura-se obter o conhecimento de determinados atributos de interesse, através de valores estimados pelo modelo criado.

De acordo com a actividade que se pretende desenvolver e com o objectivo definido para a aplicação da DCBD, define-se a tarefa a desenvolver, escolhem-se as técnicas de DM para a sua concretização e executam-se as fases para construção do modelo.

2.1. Tarefas de Data Mining

Os objectivos a atingir com a DCBD enquadram-se numa ou mais tarefas de DM, podendo estas ser de **classificação, segmentação, associação, sequenciação** ou de **sumariação.**

As actividades de **classificação** geram um conjunto de regras a partir da análise dos dados. Para tal, definem-se classes e classificam-se os dados de um conjunto de treino, a partir do qual se geram as regras. Essas regras são utilizadas no futuro para classificar novos dados, identificando de forma automática a classe a que pertencem.

A **segmentação**, também conhecida por ***clustering***, é uma actividade idêntica à classificação, mas onde as classes não estão predefinidas. Nesta actividade são identificadas as classes, de acordo com uma análise automática dos dados e que servirão para proceder à sua classificação.

As Base de Dados armazenam valores para atributos que, muitas vezes, estão relacionados entre si. As actividades de **associação** têm por objectivo determinar, a vários níveis de abstracção, quais os dados que podem ser relacionados, definindo um conjunto de regras de associação.

Numa análise sequencial, **sequenciação,** procuram-se padrões que identifiquem relações temporais nos dados, para transações realizadas em períodos de tempo diferentes.

A actividade de **sumariação** surge quando há necessidade de aumentar o conhecimento de uma Base de Dados de dimensionalidade elevada, descrevendo-a de forma resumida.

O objectivo deste tipo de actividades é analisar o que se passa nas Bases de Dados de forma a descrever os dados e evidenciar aqueles cuja análise poderá levar à descoberta de informação interessante.

2.2. Técnicas de Data Mining

As técnicas de DM consistem na aplicação de algoritmos aos dados, para detectar padrões válidos. Estas técnicas combinam aptidões de diferentes áreas de investigação, tais como bases de dados, estatística, inteligência artificial e aprendizagem automática.
A escolha dos algoritmos a utilizar no processo de DCBD depende fundamentalmente das tarefas a desenvolver, de acordo com o objectivo definido para o modelo.

Existem situações em que pelo menos duas técnicas de DM são combinadas, de acordo com as tarefas a realizar, procurando obter-se resultados com o máximo grau de confiança. A escolha e forma de combinação destes algoritmos é um processo iterativo, sendo repetido tantas vezes quantas as necessárias, em função da análise de resultados obtidos e das reformulações que esses resultados sugerem.
Embora existam várias técnicas de DM, estas podem ser agrupadas em quatro grandes categorias: **Redes Neuronais**, **Indução de Regras, Algoritmos Genéticos** e **Aproximação de Vizinhanças** [Santos 2001].

2.2.1. Redes Neuronais

As Redes neuronais são modelos muito simples que simulam o funcionamento do sistema nervoso humano. A partir de um conjunto de elementos (nodos), organizados em camadas e ligados entre si por neurónios, a rede vai propagando os valores dos nodos, alterando-os através da atribuição de pesos aos neurónios [Berry e Linoff 2000]. Este processo é repetido várias vezes e os pesos atribuídos vão sendo ajustados, em função da aprendizagem obtida em cada iteração.
Os modelos encontrados são normalmente utilizados em tarefas de classificação e segmentação.

As críticas apontadas a estes algoritmos relacionam-se com a falta de transparência do processo de decisão dentro da rede e nas dificuldades sentidas para interpretação do significado dos valores simbólicos associados aos pesos. Por isso, são mais utilizados quando os resultados são mais importantes do que o entendimento sobre como funciona o modelo e dos critérios que fundamentam as decisões [Berry e Linoff 2000].

2.2.2. Indução de Regras

Os algoritmos de indução de regras permitem gerar árvores de decisão ou regras de associação.

As árvores de decisão apresentam-se como estruturas em árvore, representando uma série de regras que apontam para uma classe ou valor. Cada *nodo* da árvore representa um atributo, a

cada ramo está associado um valor possível para esse atributo e as folhas da árvore representam as classes, isto é, as decisões possíveis [Santos 2001].

As regras de associação identificam relacionamentos entre os dados, apresentando-os numa linguagem natural, facilmente explicada e compreendida pelos utilizadores [Berry e Linoff 2000].

Este tipo de algoritmos é normalmente utilizado em tarefas de classificação, associação, sequenciação e sumariação. Por serem bastante explícitos relativamente à detecção de tendências nos dados, são ainda uma boa escolha quando se pretende seleccionar os atributos mais importantes, para definir as entradas de uma rede neuronal.

2.2.3. Algoritmos genéticos

Os algoritmos genéticos foram desenvolvidos para que as informações referentes a um determinado sistema pudessem ser codificadas de maneira análoga aos cromossomas biológicos, apresentando semelhanças com o processo evolutivo das espécies. Os algoritmos iniciam-se com um conjunto de regras, que vão sendo apuradas através da sua submissão a operadores de selecção e reprodução [Santos 2001].

São técnicas de DM normalmente utilizadas em tarefas de classificação e sumariação.

2.2.4. Aproximação de vizinhanças

Os algoritmos de aproximação de vizinhanças estão baseados no princípio de que registos semelhantes estão próximos, quando analisados numa perspectiva espacial [Santos 2001]. Cada região identificada pela proximidade de registos, interpretados como pontos no espaço, define uma classe, com características comuns aos registos que representa.

São técnicas utilizadas em actividades de segmentação ou de sumariação.

2.3. Fases do processo de DCBD

Para a aplicação das técnicas de DM apresentadas é necessário proceder a tarefas de selecção, tratamento e pré-processamento dos dados, sem as quais a DCBD não será bem sucedida.

Assim, no processo de DCBD realizam-se várias tarefas, dedicando-se as primeiras à preparação dos dados, sobre os quais se vão aplicar técnicas de DM, após o que se segue a validação e incorporação de conhecimento já existente. Todo o processo é muito iterativo, pelo que no final de cada tarefa pode existir a necessidade de voltar a realizar tarefas anteriores, para incluir alterações identificadas em fases mais avançadas.

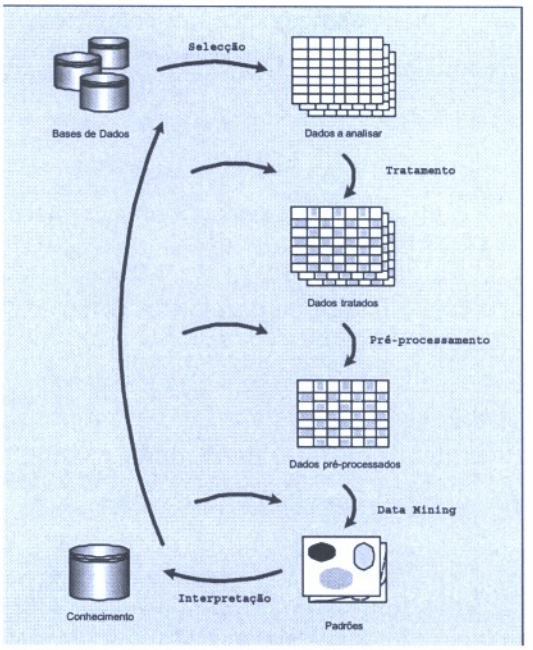


Figura 1 – Fases do processo de DCBD (Adaptado de FAYYAD et al. 1996)

Na figura 1 apresentam-se as 5 etapas que integram o processo de descoberta de conhecimento:

- Selecção dos dados;
- Tratamento dos dados;
- Pré-processamento dos dados;
- Aplicação de algoritmos de Data Mining;
- Interpretação e validação dos resultados.

Na **selecção dos dados** verificam-se os dados disponíveis e excluem-se os que não são relevantes para o processo de descoberta de conhecimento, nomeadamente aqueles que possuem carácter meramente informativo.

No **tratamento dos dados** realizam-se as operações de limpeza dos mesmos, eliminando-se registos em duplicado, eventuais erros de digitação e outras incorrecções detectadas mediante a verificação de inconsistências.

Concluídas as duas fases anteriores é feito um **pré-processamento dos dados**, de forma a facilitar a sua análise. Atributos com valores contínuos são transformados em valores discretos, para reduzir o número de linhas distintas. Os dados passam a ser agrupados e analisados por classes.

Esta tarefa é fundamental no processo de descoberta de conhecimento e raramente termina na primeira iteração. Pelo contrário, esta fase é normalmente repetida várias vezes a fim de melhorar os resultados.

Na fase seguinte procede-se à análise dos dados resultantes do pré-processamento, aplicando os

algoritmos de data mining. O processo raramente fica completo apenas pela aplicação de um único algoritmo, pelo que normalmente se combinam dois ou mais, de acordo com as tarefas a realizar.

Numa última etapa, analisam-se os dados obtidos nas fases anteriores, aplicando-se os modelos encontrados a novos conjuntos de dados, para avaliar o seu desempenho, perante dados que lhe são desconhecidos.

É nesta fase que se avalia o conhecimento produzido e se determina a validade dos resultados gerados pelos algoritmos de DM.

3. Descoberta de conhecimento na Base de Dados de Sítios Arqueológicos de Trás-os-Montes

O povoamento dos territórios é feito de acordo com variáveis de ambiente e factores económicos, sociais e políticos. A estas variáveis são atribuídas diferentes ponderações de acordo com a maior ou menor importância que é atribuída aos factores funcionais ou culturais.

Durante o período da proto-história o povoamento de Trás-os-Montes valorizava muito a questão estratégica e defensiva, aproveitando os recursos naturais do território para posicionar os seus habitats. A sua economia assentava em actividades agro-silvo-pastoris, encontrando-se nos diferentes povoados recursos diferenciados capazes de proporcionar uma autarcia económica [Lemos 1993].

Já a matriz de povoamento Romana não é tanto influenciada por factores ambientais como na proto-história. Conjugados com o quadro ambiental articularam-se ainda outros factores, destacando-se os de ordem político-administrativa e cultural.

Considerando as características diferenciadas detectadas na estratégia de povoamento do período da Proto-história e do Romano e uma vez que a maioria dos arqueossítios da base de dados de património se reportam a estas duas cronologias, vamos dar mais enfoque a estas duas cronologias no modelo preditivo a desenvolver.

Este trabalho assenta num conjunto de informação sobre o património arqueológico de Trás-os-Montes, reunida num catálogo elaborado por F. S. Lemos [Lemos 1993]. A esta base de dados foram acrescentados dados relativos a sítios arqueológicos registados em trabalhos efectuados pela Unidade de Arqueologia da Universidade do Minho, no âmbito de um projecto de elaboração dos Planos Directores Municipais da região de Trás-os-Montes, coordenado pela CCRN (Comissão de Coordenação da Região Norte).

Todos estes dados foram seleccionados, tratados e pré-processados, para serem sujeitos a técnicas de DM, com vista a elaboração de um modelo preditivo. Qualquer aplicação das técnicas de DM directamente sobre os dados, sem serem submetidos a uma selecção, tratamento e pré-processamento cuidada e criteriosa, conduzirá a resultados confrangedores, quando se pretende a obtenção de conhecimento válido.

Grande parte das Bases de Dados de áreas como a Arqueologia, Ciências Sociais ou Medicina, foram organizadas e preenchidas com objectivos diversos, que não a descoberta de conhecimento. Deste modo, estas bases de dados apresentam-se de um modo geral muito incompletas, com dados vagos e imprecisos [Rodrigues et al. 1998]. Neste contexto, as tarefas de limpeza dos dados e de transformação assumem um relevo muito especial e consomem grande parte dos recursos temporais para construir modelos válidos e úteis.

Após a realização destas tarefas sobre os dados e, tendo em mente o objectivo do trabalho de prever a localização de sítios proto-históricos e Romanos, foram aplicadas duas técnicas de DM, as Redes Neuronais e as Árvores de Decisão. Estas técnicas foram utilizadas por serem recomendadas em tarefas de previsão [Berry e Linoff 2000].

3.1. Selecção dos dados

A informação de base disponível para este trabalho está representada nas tabelas da Figura 2 e a ferramenta de descoberta de conhecimento escolhida foi o *Clementine v5.2 do SPSS Inc.*
Sobre as tabelas de dados iniciais foi feita uma selecção, que constituiu a primeira fase do processo de DCBD. Tendo em mente o objectivo do trabalho, começou-se por retirar todos os dados que, por terem carácter meramente informativo, não são relevantes para este processo. Foi o caso das **Referências Bibliográficas**, dos **Topónimos** dos sítios, do **Número** atribuído no catálogo, do código de divisão administrativa (**LOCADM**), do **Lugar** e do número da **Carta Militar**.

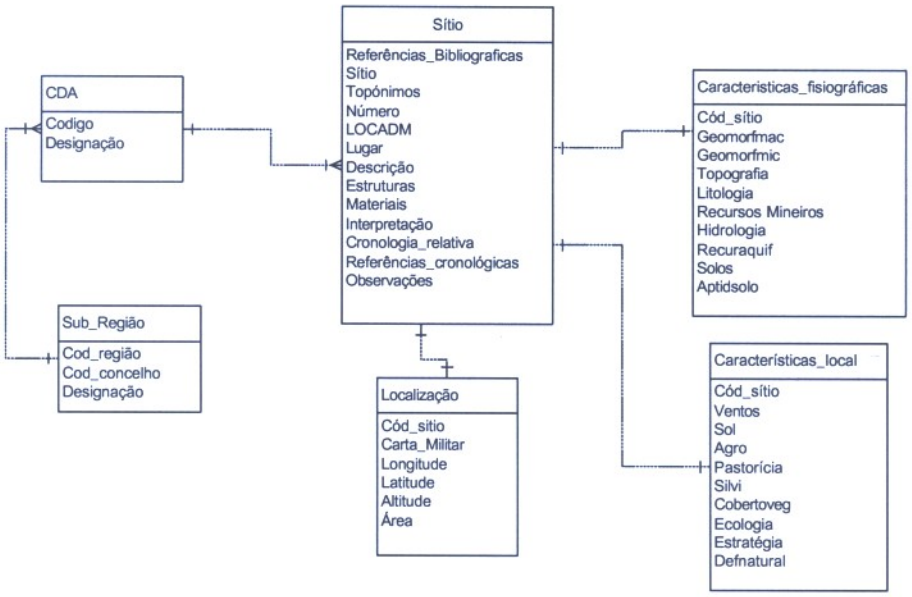


Figura 2 – Componentes da Base de Dados de sítios arqueológicos [Lemos 1993]

Aplicada a **selecção dos dados**, mediante o tipo de informação contida nos diversos atributos, procedeu-se a uma avaliação dos valores dos campos seleccionados.

3.2. Tratamento dos dados

O tratamento dos dados seleccionados na fase descrita anteriormente iniciou-se com o processo de limpeza dos mesmos, tendo sido dada especial atenção ao tratamento dos valores omissos e dados inconsistentes ou inválidos.

Os dados disponíveis para este trabalho, resultaram da integração de várias bases de dados, onde os objectivos que nortearam a sua construção eram distintos. Este factor fez com que dados importantes para uma Base de Dados não se revelassem imprescindíveis no preenchimento da informação de outras Base de Dados. Tal dualidade de critérios terá contribuído para a elevada taxa de valores omissos verificada para alguns campos.

Sempre que possível foram preenchidos os campos nulos. Foi o caso de valores omissos para a **altitude**, a **geomorfologia**, a **topografia** e **tipos de solos**, que foram preenchidos quando eram identificáveis pelas as respectivas coordenadas geográficas. Para o efeito utilizaram-se cartas topográficas que possibilitaram a recolha dessa informação.

Para alguns dados como as características do local ao nível da **Exposição solar**, **Ventos**, **Agricultura**, **Pastorícia**, **Silvicultura**, **Cobertura vegetal**, **Recursos mineiros**, **Estratégia** e **Defesa natural**, a percentagem de valores em falta era superior a 50%, pelo que, não sendo possível preencher os respectivos dados, se optou por retirá-los do conjunto de dados a analisar.

Nestas condições e dado o tipo de informação a tratar, não fazia aqui sentido atribuir-lhes valores prováveis, esperados ou mesmo gerados, dado a elevada taxa de ruído que estaria a ser introduzida no sistema.

Utilizando ferramentas de manuseamento de dados, foram encontrados e retirados todos os registos com informação em duplicado e que resultaram da junção de Bases de Dados distintas, onde alguns sítios arqueológicos foram objecto de múltiplos tratamentos.

Recorrendo aos gráficos disponibilizados pelo *Clementine*, como os histogramas e gráficos de distribuição foi possível visualizar os dados e identificar algumas inconsistências e erros de digitação. Todos os dados deste tipo foram identificados e foram corrigidos ou eliminados, nos casos em que não era possível encontrar com exactidão o respectivo valor.

Os histogramas realizados sobre os valores de entrada permitiram também identificar alguns valores isolados. Foi feita uma avaliação registo a registo de forma a distinguir casos isolados de inconsistências ou de erros de digitação. Valores isolados como "alvéolo" para a coluna **Topografia**, "Paleolítico superior" para **Cronologia** ou ainda os "Vertissolos" e "Antropossolos" para o **Tipo de solos**, foram identificados.

De acordo com o conhecimento arqueológico já existente decidiu-se manter ou excluir cada valor isolado, de acordo com a sua criticidade. Sempre que este tipo de dados não se revelava um elemento fundamental para a construção do modelo ou por não existirem exemplos suficientes para o modelar, procedeu-se à sua remoção para simplificação do modelo.

Para os dados relativos à **Topografia** e **Cronologia** considerou-se que, não sendo valores críticos, eles não seriam incluídos. No caso dos dados relativos ao **Tipo de solos** considerou-se a sua inclusão, por serem tipos de solos relevantes na região de TM e serem representativos de um aproveitamento e utilização característico da região.

3.3. Pré-processamento dos dados

Terminada a fase de **tratamento de dados** procedeu-se ao seu **pré-processamento**, onde os dados são transformados na sua forma final antes de serem analisados pelos algoritmos de DM.

As principais transformações operadas sobre os dados de sítios arqueológicos centraram-se na eliminação de variáveis correlacionadas e na análise e normalização de variáveis com significância no modelo final.

Algumas colunas da base de dados como a **Hierarquia hidrográfica** e a **Tipologia** foram obtidas pela generalização dos dados e criando uma hierarquia de conceitos para que as regras

geradas pelos algoritmos de DM possam ser mais simples, claras, concisas e facilmente generalizáveis.

3.3.1. Normalização dos dados

Valores contínuos da **Altitude**, **Longitude** e **Latitude** foram transformados em valores discretos, para serem analisados e agrupados por classes.

Para a **Altitude** foram criadas as classes representadas na Figura 3, de acordo com o mapa do esboço geomorfológicos de TM [Ribeiro et al. 1987].

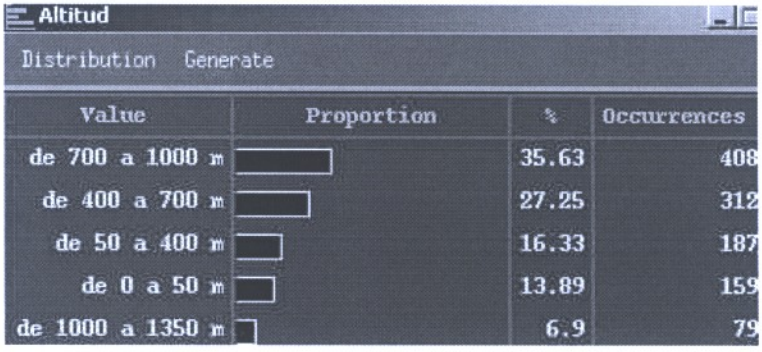


Figura 3– Classes criadas para os valores da **Altitude**

Para a **Longitude** e **Latitude** foram usados critérios de classificação centrados na distribuição uniforme dos dados.

O campo **Tipologia** resultou também de um trabalho de normalização dos dados da coluna **Interpretação**. Designativos como *"Vicus"*, "habitat" e "povoados" representam, para o objectivo deste trabalho, o mesmo conceito e foi-lhes atribuída a designação única de "povoado".

O campo da **Cronologia relativa** contém as referências a datas atribuídas aos sítios arqueológicos. Também estes dados foram normalizados e foram adoptadas classes cronológicas que uniformizam e agregam os sítios por períodos temporais.

Durante a fase de pré-processamento dos dados verificou-se ainda que alguns registos continham múltiplos valores para **Estruturas**, **Interpretação** e **Cronologias**.

A título de exemplo refere-se o caso de registos com dados como "vicus; necrópole", ou "tesouro monetário; habitat romano" que, aparecendo na coluna **Interpretação**, correspondem cada um a duas classes de **Tipologias**. Fazendo a correspondência com as classes criadas para agrupamento e normalização dos dados, os valores de **Interpretação** "*Vicus*; necrópole"

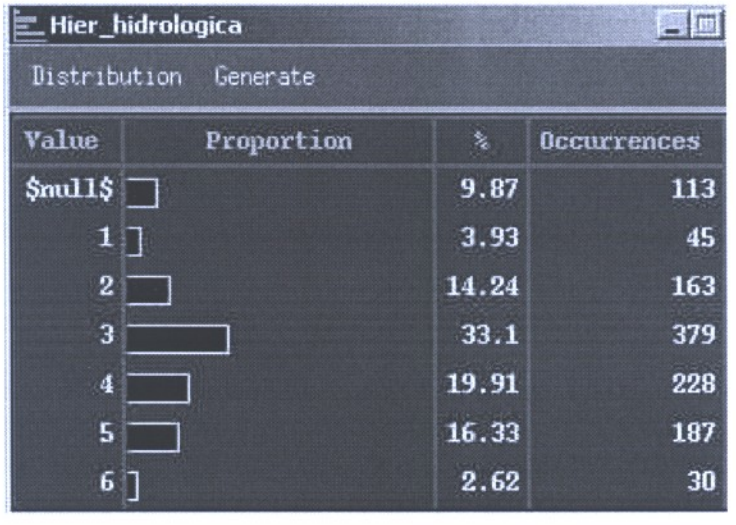
correspondem às **Tipologias** "povoado" e "necrópole" e os valores como "tesouro monetário; habitat romano" correspondem às **Tipologias** "tesouro" e "povoado".

Da mesma forma, encontrou-se muitas vezes para o mesmo local, um registo de um sítio arqueológico correspondente a duas épocas distintas. Acontece que o local escolhido para localizar um habitat da "Idade do Ferro", pode ter perdurado na "época Romana". Assim, desmembrou-se este sítio em dois, dada a sua correspondência a duas épocas cronológicas distintas.

3.3.2. Generalização dos dados

A hidrografia é caracterizada na tabela de dados pelas colunas **Hidrologia** e **Recursos aquíferos**. Para facilitar o tratamento e análise desta informação, criou-se uma nova coluna – **Hierarquia hidrográfica**, que generaliza os dados relativos à **Hidrologia** e **Recursos aquíferos**. A partir do nome do rio, ribeiro ou linha de água que está próximo do sítio arqueológico, atribuiu-se um valor de 1 a 6, os quais representam os níveis na hierarquia da bacia hidrográfica. O nível 1 está associado a locais próximos dos cursos de água principais, crescendo este valor na razão directa da estrutura hidrográfica. No caso em estudo foram considerados de nível 1 os locais próximos dos rios Douro e Cávado, por serem os dois cursos de água principais que terminam no oceano.

Do gráfico representado na Figura 4 será interessante realçar o facto de quase 70% dos sítios arqueológicos se posicionarem nos níveis hidrológicos 3, 4 e 5, ou seja, nos níveis intermédios da rede hidrográfica.



Value	Proportion	%	Occurrences
$null$		9.87	113
1		3.93	45
2		14.24	163
3		33.1	379
4		19.91	228
5		16.33	187
6		2.62	30

Figura 4– Classes criadas para a **Hierarquia hidrológica**

Também o campo **Tipologia** resultou do pré-processamento realizado aos campos **Interpretação**, **Estruturas** e **Materiais** que existiam na Base de Dados inicial. Estes campos contêm um pequeno texto com a descrição do tipo de sítio identificado e dos materiais lá encontrados. A cada descritivo fez-se corresponder uma ou mais classes criadas para o efeito e foi registada a equivalência no campo **Tipologia**.

As classes de **Tipologias** criadas são os "povoados", "fortificações", "povoados fortificados", "necrópoles", "santuários", "redes viárias", "arte rupestre", "epigrafia", "minas", "esconderijos", "tesouros" e "vestígios".

Numa breve análise da Figura 5 constata-se que a base de dados contém maioritariamente arqueossítios do tipo "necrópoles", "povoados" e "povoados fortificados". Poderá no entanto causar alguma surpresa o facto de existirem mais "necrópoles" do que "povoados". Tal acontece porque do "Período megalítico" foram identificadas várias "necrópoles megalíticas" e foram reduzidas as localizações de "povoados", associados a esse período temporal.

Tipologia

Distribution Generate

Value	Proportion	%	Occurrences
Epigrafia		1.75	20
Fortificação		2.27	26
Minas		0.44	5
Necrópole		33.54	384
Povoado		19.04	218
Povoado fortificado		22.97	263
Rede viária		6.29	72
Santuário		4.02	46
Tesouro		1.75	20
Vestígios		5.33	61

Figura 5 – Distribuição das **Tipologias**

3.3.3. Eliminação de variáveis correlacionadas

A Bases de Dados de trabalho, representada na Figura 2, contém alguns campos com informação correlacionada e que deve ser retirada. É o caso da coluna **Referências cronológicas** que, para o objectivo de DCBD, introduz informação correlacionada com a coluna **Cronologia**, pelo que foi suprimida.

As colunas **Hidrologia** e **Recursos aquíferos** apresentam também informação altamente correlacionada com a coluna **Hierarquia hidrográfica**, criada para generalização dos dados que

caracterizam a rede hidrográfica. Deste modo, as colunas que lhe deram origem, **Hidrologia** e **Recursos aquíferos** foram retiradas.

Retiraram-se também as colunas **Interpretação, Estruturas** e **Materiais** por estarem altamente relacionadas com os dados da **Tipologia**, que resultaram da sua normalização e generalização.

3.3.4. Tabela de dados a tratar

A Tabela 1 representa o conjunto de dados resultante da selecção, tratamento e pré-processamento do conjunto inicial.

Sítios arqueológicos
Tipologia
Cronologia
Latitude
Longitude
Altitude
Geomorfologia
Geomorfologia_mic
Topografia
Litologia
Hierarquia_hidrográfica
Solos
Paisagem

Tabela 1 – Tabela de dados após pré-processamento

3.3.5. Análise de relações entre os dados

A partir dos dados da Tabela 1 iniciou-se a fase de exploração dos dados, utilizando algumas das técnicas de visualização disponíveis no *Clementine*, nomeadamente os *Web Nodes*. Esta forma de apresentação gráfica permite uma fácil interpretação dos relacionamentos existentes entre os dados e possibilita a identificação das variáveis mais influentes e que afectam as variáveis de saída do modelo, neste caso a **Tipologia** dos sítios.

Estes gráficos permitem identificar algumas relações interessantes entre dois ou mais atributos simbólicos. As ligações são expressas graficamente, através de pontos, linhas e linhas sombreadas. As relações mais fortes são as desenhadas a traço contínuo mais carregado, passando a tracejada quando estamos perante relações fracas. Dados não ligados indicam que não foi identificada qualquer relação entre eles.

Fazendo uma análise ao gráfico da Figura 6 verifica-se que os "povoados" se distribuem por todo o território, não havendo contextos preteridos. Apesar da região de TM ter características geomorfológicas bastante diversificadas e dos recursos estarem dispersos, a Proto-história introduziu inovações tecnológicas, depois melhoradas e aperfeiçoadas no período "Romano", que potenciaram um aproveitamento mais equilibrado dos recursos disponíveis, permitindo uma ocupação de territórios mais abrangente [Lemos 1993].

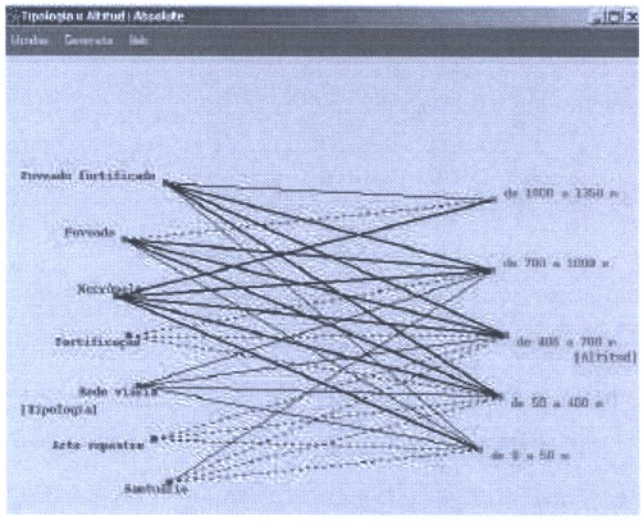


Figura 6 – *Web node* que relaciona a **Tipologia** com a **Altitude**

Curiosamente as "fortificações" aparecem a altitudes intermédias e os "povoados fortificados" estão distribuídos por todo o território. De facto, as "fortificações" isoladas ou associadas a "povoamentos" não privilegiam apenas a "altitude", podendo concluir-se que as "fortificações" tinham um carácter multifuncional, servindo para:

- Controle e delimitação do território envolvente;
- Estratégias de defesa;
- Aspectos simbólicos, funcionando como valor arquitectónico;
- Delimitação material dos povoados.

Por outro lado a rede viária não se distribui por todos os patamares de "altitude", o que confirma o conhecimento de que os engenheiros romanos evitavam não só as cotas muito elevadas, como também as grandes variações. O facto de haver "povoados" e "povoados fortificados" acima dos 1000 m, sem que a "rede viária" seja localizada nesse patamar de "altitude" deve-se ao traçado das vias não acompanhar a distribuição dos "povoados", mas obedecer a um plano lógico e a um estudo prévio da geomorfologia dos terrenos.

Todas as outras **Tipologias** se localizam preferencialmente a cotas mais baixas, nomeadamente a "arte rupestre" que, nesta região, se encontra encaixada no fundo dos vales [Lemos 1993].

A Figura 7 introduz uma nova variável, em relação ao gráfico da Figura 6, que é a **Cronologia**, analisando a distribuição por **Altitudes** dos arqueossítios da "Idade do Ferro" e do período "Romano".

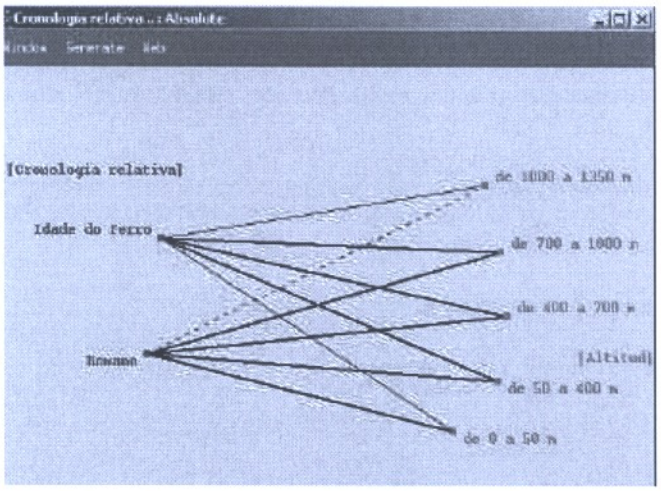


Figura 7 – *Web node* que relaciona a **Cronologia** dos sítios com a **Altitude**

Embora já o gráfico da Figura 6 indicasse uma ocupação abrangente do território de TM, pode-se agora constatar que essa abrangência era maior no período da "Idade do Ferro" do que no período "Romano" que se lhe seguiu.

Tal poderá dever-se ao facto de na "Idade do Ferro" o povoamento estar distribuído em função de um equilíbrio com o contexto natural e, no período "Romano", haver uma hierarquia de povoamento influenciada por uma rede de caminhos e por uma nova economia [Lemos 1993].

O gráfico da Fgura 8 estabelece as relações entre as **Tipologias** e a **Hierarquia hidrográfica** e, tal como já se tinha verificado no gráfico da Figura 4, os locais mais próximos e mais afastados dos cursos de água principais são os menos ocupados pelos "povoados", "fortificações" e "povoados fortificados".

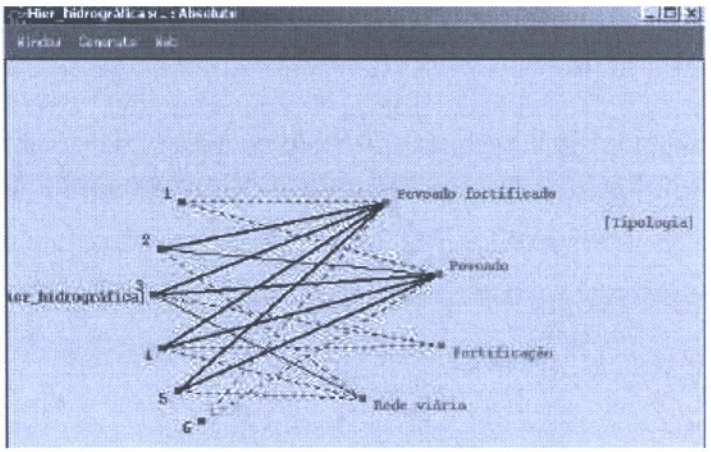


Figura 8 – *Web node* que relaciona as **Tipologias** com a **Hierarquia hidrográfica**

O facto dos locais mais próximos do curso de água primário da região, que para a maioria dos sítios é o rio Douro, se caracterizarem por terem poucas **Tipologias** localizadas poderá ter várias explicações. Uma delas pode atribuir-se à circunstância deste rio ter margens com vertentes escarpadas e de difícil acessibilidade, factor que poderia ter impedido uma ocupação mais extensiva destes locais. Outra justificação está relacionada com a prospecção pouco intensiva destas áreas, no âmbito de projectos de Arqueologia.

3.4. Aplicação de algoritmos de Data Mining

Terminadas as fases de compreensão dos dados, selecção, tratamento e pré-processamento dos mesmos, procedeu-se à aplicação de técnicas de modelação avançada.

A Fgura 12 ilustra a primeira fase da aplicação de técnicas de DM, em que o conjunto de dados resultante das fases anteriores (**TMO**) é subdividido em dois grupos, a que chamamos de **treino** e **testes**. É sobre o primeiro conjunto, mais pequeno, designado de treino, que irá ser gerado o modelo. Os testes e a avaliação do desempenho do modelo são feitos por aplicação deste ao conjunto de testes, cujos dados lhe são desconhecidos.

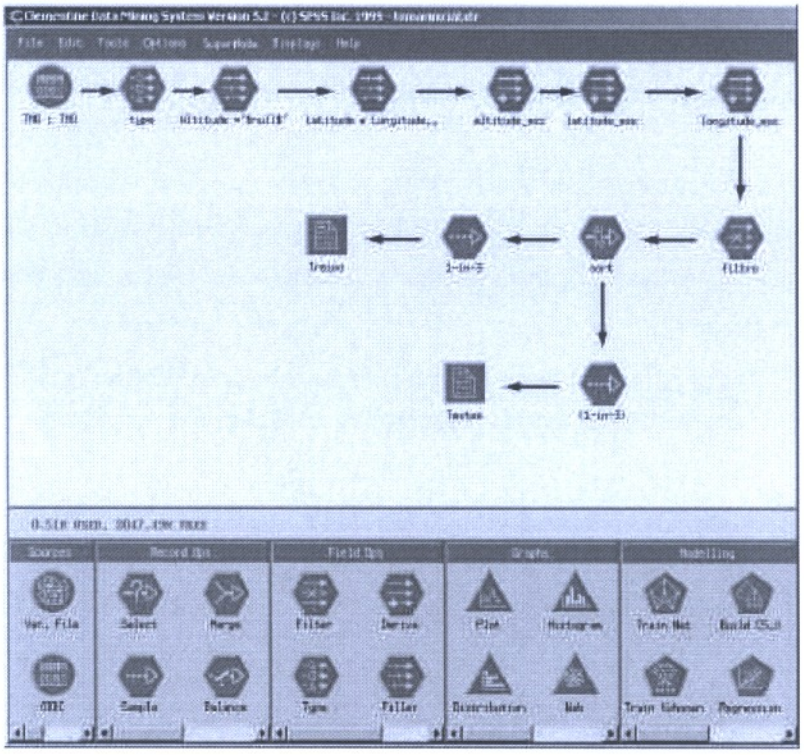


Figura 9 – Detecção de sítios arqueológicos – conjunto de treino e de testes

Dado o carácter previsional do modelo que se pretende construir aplica-se ao conjunto de treino um algoritmo de indução de árvores de decisão e posteriormente um algoritmo de redes neuronais.

Com o objectivo de determinar o conjunto de atributos relevantes para a inferência de sítios arqueológicos em TM, aplicou-se ao ficheiro de treino o algoritmo C5.0. Como resultado obteve-se um modelo (Figura 10), do qual se apresentam algumas regras onde estão identificados os atributos relevantes para a caracterização dos sítios arqueológicos e da sua **Tipologia**, para o período "Romano".

```
Cronologia relativa Romano
    Geomorfmic [Alto Chã Cume] -> Epigrafia
    Geomorfmic Alvéolo -> Povoado
    Geomorfmic Arriba -> Vestígios
    Geomorfmic Castelo -> Vestígios
    Geomorfmic Inselberg -> Povoado
    Geomorfmic Ladeira -> Rede viária
    Geomorfmic Lombeiro -> Povoado
    Geomorfmic Margem -> Povoado
    Geomorfmic Monte -> Minas
    Geomorfmic Outeiro -> Povoado
    Geomorfmic Terraço -> Povoado
    Geomorfmic Vertente -> Rede viária
    Geomorfmic Esporão
        Altitud ['de 0 a 50 m' 'de 1000 a 1350 m'] -> Santuário
        Altitud de 50 a 450 m -> Santuário
        Altitud de 700 a 1000 m -> Fortificação
        Altitud de 450 a 700 m
            Hier_hidrográfica [$null$ 1 3 4 6] -> Povoado
            Hier_hidrográfica 2 -> Povoado
            Hier_hidrográfica 5 -> Santuário
    Geomorfmic Planalto
    Geomorfmic Vale
    Geomorfmic $null$
    Geomorfmic Cabeço
```

Figura 10 – Algumas regras obtidas pela aplicação do algoritmo C5.0

Os atributos que, no modelo obtido com o algoritmo C5.0, não são apresentados como relevantes são filtrados no conjunto de treino, para posteriormente serem submetidos a um algoritmo de rede neuronal.

Os resultados preliminares obtidos com o treino de uma rede neuronal não se apresentaram muito satisfatórios, pelo que foram feitas algumas experiências, ao nível do tipo de rede neuronal utilizada e da distribuição quantitativa dos registos de treino e testes. Dado o número reduzido de registos inicial, foi ainda realizado o balanceamento para o campo **Tipologias**.

As alterações realizadas produziram resultados mais satisfatórios, cuja análise qualitativa da aplicação ao conjunto de testes se apresenta na Figura 11. Estes resultados foram encontrados submetendo os dados resultantes da análise das regras obtidas pelo modelo C5.0, a uma rede neuronal que apresentava uma *Predicted Accuracy* de 88,47%. O atributo **$C-Tipologia** representa a saída da Árvore de Decisão, enquanto que o campo **$N-Tipologia** diz respeito à previsão realizada pela Rede Neuronal gerada.

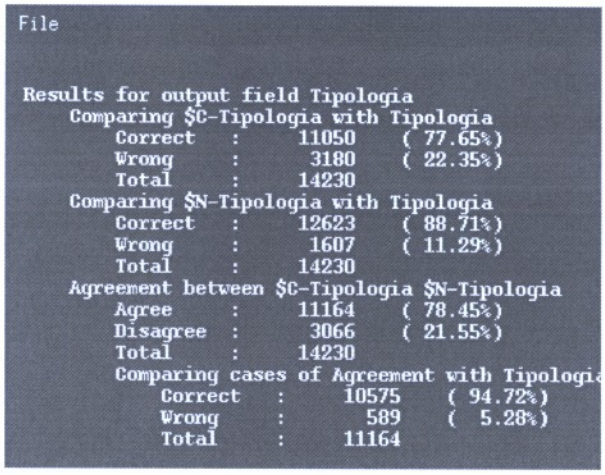


Figura 11 – Análise qualitativa da aplicação do modelo ao conjunto de testes.

4. Avaliação de Resultados

O *focus* deste trabalho foi dado à aplicação do processo de Descoberta de Conhecimento a uma Base de Dados de sítios arqueológicos. Sendo um pressuposto assumido para o realizar, que as fases de tratamento e pré-processamento dos dados são fundamentais para aplicação com sucesso das técnicas de DM, constatou-se que para o tipo de bases de dados utilizadas, a criticidade destas fases é muito elevada. As Bases de Dados de Arqueologia, bem como as de Medicina ou das Ciências Sociais, têm normalmente informação por vezes vaga, muitas vezes omissa, com contextos subjectivos e níveis de agregação diferenciados.

A incorporação de conhecimento arqueológico existente foi fundamental na fase de pré-processamento dos dados, bem como na avaliação do conhecimento produzido nos modelos encontrados. Sem esta componente pode-se estar a trabalhar em pressupostos errados e a encontrar conhecimento que pode não ser útil ou válido em Arqueologia.

No entanto, com um trabalho intensivo de selecção de dados, normalização e generalização, conseguiram-se resultados que podem ser indicadores muito interessantes e úteis. Os modelos obtidos sobre a ocupação de território, em Trás-os-montes, para a Idade do Ferro e período "Romano", poderão desde já contribuir para a elaboração de cartas de risco, com vista a protecção do património arqueológico.

5. Conclusão e trabalho futuro

Este trabalho foi desenvolvido com o objectivo de construir um modelo preditivo, que facilite a identificação de sítios arqueológicos.

Para concretizar este objectivo, integrou-se o conhecimento arqueológico com a Descoberta de Conhecimento em base de dados e obteve-se um modelo preditivo de Património Arqueológico, em função de determinados parâmetros e que fornecem indicadores sobre a localização e tipologia dos arqueossítios, na região de Trás-os-Montes.

Estes modelos preditivos de património arqueológico poderão ter diversas aplicações.

No âmbito da Arqueologia poderá ser um valioso instrumento ao serviço de estudos prospectivos. Em Arqueologia preventiva recorrem-se a indicadores vários que têm a ver com bibliografia e com o reconhecimento do terreno, para detecção de vestígios diversos que indiciem a existência de património arqueológico.

Poder recorrer a um modelo que dê ao arqueólogo indicações sobre as áreas de maior probabilidade de encontrar Património, constituirá desde logo um instrumento válido e útil para a Arqueologia.

No âmbito da administração de território e valorização do património estes modelos poderão ser preciosos auxiliares na delimitação de áreas, onde há fortes probabilidades de existir arqueossítios ocultos e que, consertadas com a cartografia de Património registado e inventariado, poderão constituir um instrumento válido e útil para a gestão do Património.

Este modelo poderá ser ainda melhorado quer pela alteração de alguns pressupostos, resultantes de indicadores fornecidos pela análise aos resultados já obtidos, quer pela inclusão de novos dados, entretanto recolhidos sobre o património da região estudada.

De acordo com a utilização a dar a estes modelos, no âmbito da Arqueologia ou no âmbito da gestão de património, deverá ser desenvolvido um sistema específico de interface e de visualização de resultados.

A mais valia deste trabalho advém do facto de conjugar o saber em Arqueologia com o saber da área dos Sistemas de Informação e Descoberta de Conhecimento, para conseguir resultados cuja aplicação e utilidade transcende as duas áreas que lhe deram origem.

A possibilidade de evidenciar a distribuição no espaço e no tempo de diferentes tipos de arqueossítios, poderá ser um outro contributo para aprofundar o conhecimento das sociedades do passado e das suas estratégias territoriais.

O desenvolvimento de interfaces com ambientes de Sistemas de Informação Geográfica, ambientes multimédia e até ambientes virtuais, poderá tornar este sistema ainda mais atractivo e

fundamentalmente mais útil à comunidade científica e ao público em geral.

6. Referências Bibliográficas

Berry, Michael e Gordon Linoff, *"Mastering Data Mining – The Art and Science of Customer Relationship Management"*, Wiley Computer Publishing, New York, USA, 2000.

Daveau, Suzanne, *"Espaço e Tempo: evolução do ambiente geográfico de Portugal ao longo dos tempos pré-históricos"*, Clio, 2, 1980.

Fayyad, U. M., G. Piatetsky-Shapiro, e R. Uthurusamy (Eds.), *"Advances in Knoledge Discovery and Data Mining"*. The MIT Press, Massachusetts, 1996.

Fayyad, Usama, Georges Grinstein e Andreas Wierse, *"Information Visualization in Data Mining and Knowledge Discovery"*, 2001.

Han, Jiawei e Micheline Kamber, *"Data Mining: Concepts and Techniques"*, Morgan Kaufmann Publishers, 2001.

Lemos, Francisco Sande, *"Povoamento Romano de Trás-os-Montes Oriental"*, Universidade do Minho, 1993, Tese de Doutoramento.

Santos, Maribel Yasmina Campos Alves, *"PADRÂO – Um sistema de Descoberta de Conhecimento em Bases de Dados Georeferenciadas"*, Universidade do Minho, 2001, Tese de Doutoramento.

Ribeiro, Orlando e Hermann Lautensach, *"Geografia de Portugal"*, Vols. I,II,III e IV, Edições João Sá da Costa, Lisboa, 1987.

Rodrigues, Maria de Fátima, Carlos Ramos e Pedro Rangel Henriques, *"Extracção de Conhecimento em Sistemas de Informação Imprecisos"*, EEI'98, 1998.

SPSS, Clementine, User Guide,Versão 5.2, SPSS Inc., 1999.