Summary

Supervised classification problems are at heart easy to understand, but rapidly get very hard to actually solve. Many criteria can be used to determine a classifier. The quality of the classifier can only be determined afterwards by means of validation, in most cases using cross-validation and 10-fold cross-validation in particular. Several classification algorithms are popular. Based on classification power some such as Support Vector Machines and k-Nearest Neighbor can be preferred. When taking a broader range of accessibility criteria into account, more precisely: easy of use, transparency and power, it can be concluded that these powerful classification algorithms do not score very high overall. This can also explain why a classification algorithm such as C4.5 is actually more popular: it is very accessible, even if it is less powerful in the traditional sense.

Many problems that are the source of low accessibility can be attributed to the nature of the traditional data mining processes. These problems can be roughly grouped in linearity shortcomings, modularity shortcomings and inadequate user involvement shortcomings. A new process model is introduced that solves the shortcomings of the traditional processes and thus allows for more accessible classification problem solving. The feature selection and construction is now done while building the classifier. This immediately reveals the effects of these actions, making well founded decisions possible at each stage. The process is iterative, results and insights can now be immediately valorized during the following phases. By building a data model at the beginning, the user no longer has to do difficult transformations. The responsibility of being able to handle the data has been moved to the algorithms. As the process can not continue unless the user can decide how to continue, a tool following the new process is forced to incorporate the idea that information must be communicated to the user in a form the user understands. Expert knowledge from the domain expert can then be used in two ways. First, expert knowledge can be added as meta data when building the data model. Second, the domain expert can

directly influence the construction of the classifier, which can be based on expert knowledge. Tools following the new process should no longer need algorithm and parameter selections. Such a tool can display many patterns using multiple techniques, at least one of which should be acceptable. The optimal selections can be left to the user when selecting the path to continue. By communicating the patterns to the domain expert, the system does not only help the domain expert to build a classifier, but also makes the domain expert understand the data better. By doing this in an iterative process makes it less overwhelming and also confronts the domain expert with more hidden patterns. The end result is that now not only a classifier is built, but also that expert knowledge has been created. Although the new process model may not be suitable for each classification problem, due to maybe the scope or extreme properties of the problem, it does provide a model for more accessible classification problem solving in other cases.

To support this classification process, the structured classification data model is created. It defines a combination of data and a description of the structure of the data including different kinds of meta data. The structure enables the definition of attributes types, which are either numerical, nominal or ordinal in nature. The most powerful addition is the possibility to indicate whether attribute values are optional or not. It is even possible to indicate that the existence of some attribute values is dependent on some constraint. By allowing the domain expert to add this information, the preconditions are set to move the responsibility of dealing with the inherent structure of the data from the user to the classification tool. Hereby the most difficult part of preprocessing that severely limits the accessibility is addressed. Also some shortcomings associated with the lack of user involvement are remedied this way. The Structured Data Meta Classification Tree (SDMCTree) allows traditional classification algorithms to be used in combination with the structured classification data. This way the traditional classification process may also be followed, but the preprocessing is still significantly reduced. Depending on the circumstances three different algorithms to build different variations of the SDMCTree are available.

The Glass Tree classifier model complies with the aforementioned process model and uses the structured classification data model. Contrary to traditional classifiers, the Glass Tree comes in two forms. The Glass Tree Creator is used to create the structure of the classifier. The Glass Tree Classifier calibrates this structure with training data and can perform the actual classifications. The Glass Tree finds patterns in the form of candidate splits along attributes and candidate orientations for linear splits. This information is then communicated to the user with the change in classification power

they bring, the information they make available, and a visualization of the data simulating uncertainty that can be browsed through the dimensions. This information allows the user to select the path to continue, which can be growing of the tree with either a split on an attribute, a linear split, or a user guided linear split. The user can also decide to backtrack previous actions by pruning the tree. The cycle can then restart. The result is a versatile and powerful classifier that addresses the remaining accessibility shortcomings by extensively involving the user, and thereby meeting all requirements of accessibility.

Samenvatting

Supervised classification problemen zijn in essentie eenvoudig te begrijpen, maar zijn vaak moeilijk op te lossen. Vele criteria kunnen worden gebruikt om een *classifier* te bepalen. De kwaliteit van de *classifier* kan enkel later worden bepaald door middel van validatie, in de meeste gevallen *cross-validation* en *10-fold cross-validation* in het bijzonder. Verschillende classificatiealgoritmen zijn populair. Wanneer men zich baseert op classificatiekracht kunnen sommige algoritmen zoals *Support Vector Machines* en *k-Nearest Neighbor* de voorkeur wegdragen. Wanneer we een bredere waaier van toegankelijkheidscriteria in acht nemen, meer bepaald: gebruiksvriendelijkheid, transparantie en kracht, kan er worden besloten dat deze krachtige classificatiealgoritmen niet erg hoog scoren in het algemeen. Dit kan ook verklaren waarom een classificatiealgoritme als C4.5 populairder is: het is met name zeer toegankelijk, zelfs al is het minder krachtig in de traditionele betekenis.

Vele problemen die aan de bron liggen van de beperkte toegankelijkheid kunnen worden toegeschreven aan het wezen van de traditionele *data mining* processen. De problemen kunnen grofweg gegroepeerd worden in lineariteitsbeperkingen, modulariteitsbeperkingen en beperkte betrekking van de gebruiker beperkingen. Een nieuw procesmodel wordt geïntroduceerd dat deze beperkingen van de traditionele processen oplost en dus meer toegankelijk classificatieprobleem oplossen mogelijk maakt. De *feature* selectie en constructie gebeurt nu tijdens het maken van de *classifier*. Dit onthult onmiddellijk de effecten van deze handelingen, waardoor goed gefundeerde beslissingen mogelijk worden gemaakt tijdens elke fase. Het proces is iteratief, waardoor resultaten en inzichten nu onmiddellijk kunnen worden gevaloriseerd tijdens de volgende fases. Door in het begin een datamodel te bouwen, moet de gebruiker niet langer ingewikkelde transformaties uitvoeren. De verantwoordelijkheid om deze data te verwerken wordt verplaatst naar de algoritmen. Omdat het proces niet kan verdergaan behalve als de gebruiker kan beslissen hoe te vervolgen, moet een *tool* die het nieuwe proces

volgt, verplicht het concept dat de informatie naar de gebruiker moet worden gecommuniceerd invoeren. Expertkennis van de domeinexpert kan op twee manieren worden gebruikt. Ten eerste kan expertkennis worden toegevoegd als metadata tijdens het bouwen van het datamodel. Ten tweede kan de domeinexpert onmiddellijk de constructie van de *classifier* beïnvloeden, wat op basis van expertkennis kan gebeuren. *Tools* gestoeld op het nieuwe proces zouden niet langer algoritme- en parameterselectie mogen vereisen. Zo een *tool* kan vele patronen weergeven, gebruikmakend van verschillende technieken, waarvan er minstens één aanvaardbaar zou moeten zijn. De optimale keuzes kunnen aan de gebruiker worden overgelaten wanneer de manier van vervolgen wordt gekozen. Door de patronen naar de domeinexpert te communiceren, helpt het systeem niet enkel de domeinexpert met het creëren van een *classifier*, maar helpt het ook de domeinexpert de data beter te begrijpen. Door dit in een iteratief proces te doen, is dit minder overweldigend en confronteert het de domeinexpert ook met verborgen patronen. Het eindresultaat is dat er niet enkel een *classifier* wordt gecreëerd, maar ook expertkennis. Hoewel het nieuwe procesmodel misschien niet gepast is voor elke soort classificatieproblemen, door misschien de omvang of buitengewone eigenschappen van het classificatieprobleem, kan het wel fungeren als een model voor meer toegankelijk classificatieprobleem oplossen in andere gevallen.

Om het classificatieproces te ondersteunen is het gestructureerde classificatie data model gecreëerd. Het definieert een combinatie van data en een beschrijving van de structuur van de data inclusief verschillende soorten metadata. De structuur laat de definiëring van attribuuttypes toe, die numeriek, nominaal of ordinaal van aard kunnen zijn. De krachtigste toevoeging is de mogelijkheid om aan te geven of een attribuutwaarde optioneel is of niet. Het is zelfs mogelijk om aan te geven dat het bestaan van attribuutwaarden afhankelijk is van een bepaalde restrictie. Door de domeinexpert de mogelijkheid te geven deze informatie toe te voegen, zijn de voorwaarden aanwezig om de verantwoordelijkheid van het omgaan met de inherente structuur van de data te verplaatsen van de gebruiker naar de *tool*. Hierdoor is het moeilijkste gedeelte van de *preprocessing* dat ernstig de toegankelijkheid beperkt aangepakt. Ook worden andere tekortkomingen die geassocieerd kunnen worden met de te beperkte betrekking van de gebruiker hierdoor verholpen. De Structured Data Meta Classification Tree (SDMC-Tree) laat toe traditionele classificatiealgoritmen te gebruiken in combinatie met de gestructureerde classificatie data. Op deze manier kan het traditionele classificatieproces toch worden gevolgd, maar wordt de *preprocessing* wel ernstig verminderd. Afhankelijk van de omstandigheden zijn er drie ver-

schillende algoritmen om verschillende variaties van de SDMCTree te creëren beschikbaar.

Het Glass Tree *classifier* model stemt overeen met het voorgenoemde procesmodel en gebruikt het gestructureerde classificatie data model. In tegenstelling tot klassieke *classifiers*, zijn er twee vormen van de Glass Tree. De Glass Tree Creator wordt gebruikt om de structuur van de *classifier* te bepalen. De Glass Tree Classifier kalibreert deze structuur met trainingsdata en kan de feitelijke classificaties uitvoeren. De Glass Tree vindt patronen in de vorm van kandidaat splitsingen langsheen de attributen en kandidaat oriëntaties voor lineaire splitsingen. Deze informatie wordt dan gecommuniceerd naar de gebruiker samen met de veranderingen in classificatiekracht die deze met zich mee brengt, de informatie die ze beschikbaar maakt en een visualisatie van de data die de onzekerheid simuleert die doorbladerd kan worden doorheen de dimensies. Deze informatie laat de gebruiker toe de manier van vervolgen te selecteren, welke ofwel het groeien van de boom met een splitsing op een attribuut, een lineaire splitsing of een door de gebruiker gestuurde splitsing kan zijn. De gebruiker kan ook beslissen om terug te komen op vorige beslissingen door de boom te snoeien. De cyclus kan dan herbeginnen. Het resultaat is een veelzijdige en krachtige *classifier* die de resterende toegankelijkheidstekortkomingen aanpakt door intensief de gebruiker in het proces te betrekken en daardoor tegemoetkomt aan alle vereisten van toegankelijkheid.